Sequence Mining analysis on Shopping Data

MASTER'S DISSERTATION

PORTO, 2017 JANUARY

JOÃO MIGUEL DA ROCHA RIBEIRO

SUPERVISOR: DR. CARLOS SOARES

CO-SUPERVISOR: ZAFEIRIS KOKKINOGENIS

MASTER IN INFORMATICS AND COMPUTING ENGINEERING

Abstract

With easy access to information, it is only natural that people and companies try to extract the maximum value from it. For instance, every large retail brand and shopping centre in the world strives to collect data about their customers and habits. Knowledge can be extracted from this data with data mining techniques. Due to this relentless demand for new knowledge in data, learning new means of obtaining it can bring great competitive advantages.

Several works are mentioned in the literature attempting to extract knowledge from retail data. Sometimes sequence mining techniques are applied in the e-commerce context. This type of approaches tries to analyse the click-stream data as sequences of events. However, few of them have used sequence mining techniques. Here, the notion of event represents a visit a person performs to a store. These techniques analyse data representing sequences of events (e.g. clicks on web pages or visits of customers to stores), extracting frequent sub-sequences (e.g. a significant number of customers visit store A, then store C, then store B). Each visit is characterized by a number of features, such as the store name and category and the day and time of visit. One of the interesting issues to explore with sequence mining is related to the different representation of sequences that can lead to different knowledge extraction.

This dissertation reports an empirical study to extract knowledge about customer behaviour from data representing visits of customers to stores in the brick-and-mortar context using sequence mining techniques. The dataset is composed of spatial-time referenced data of customer locations in a shopping centre. A sequence is a set of one or more visits. Each visit is composed of a client ID, a store, and the time of that detection and other related information. We explore sequence mining techniques on different representations. With this study, we concluded that these techniques can be used to obtain some pieces of information contained in the data. We found evidence that they can be used to extract information in all the tested representations. We also describe other subjects and guidelines that are important to implement this solution in a retail dataset.

Resumo

Vivemos numa altura onde o acesso à informação é cada vez mais fácil. Esta facilidade leva a que pessoas e empresas tentem extrair o máximo de valor inerente. Um pouco por toda a parte as grandes marcas de retalho e de centros comerciais competem entre si para conseguirem oportunidades de acesso a dados relativos aos clientes e aos seus hábitos. Esta procura desenfreada leva a que se tentem encontrar novos meios para a sua deteção com o objetivo de se conseguir obter alguma vantagem competitiva sobre os concorrentes. Esta informação pode ser encontrada através do uso de técnicas de *Data Mining*.

Nas referências da dissertação estão presentes vários exemplos de tentativas de extração de informação no contexto de retalho. Atualmente, existem já vários exemplos de aplicação de técnicas de *sequence mining* aplicadas no retalho online. Este tipo de aplicações procura analisar a ordem de *clicks* por parte dos utilizadores em aplicações web. No entanto, as técnicas de *sequence mining* raramente são usadas. Estas técnicas analisam dados que representam sequências de eventos (por exemplo, visitas de clientes a lojas), com o objetivo de encontrar subsequências frequentes (por exemplo, um número significativo de pessoas visita em primeiro lugar a loja A, depois visita a loja C e finalmente visita a loja B). Cada visita é composta por um determinado número de características, como o nome da loja, a sua categoria e o dia e a hora em que a visita foi feita. Uma das questões interessantes a explorar com a utilização de *sequence mining* está relacionada com as diferentes representações de sequências que podem levar a diferentes extrações de conhecimento.

Esta dissertação relata um estudo empírico para a extração de conhecimento sobre o comportamento de clientes a partir de dados que representam visitas de clientes usando *sequence mining*. A base de dados é composta por dados espácio-temporais de lojas em que os clientes estiveram num centro comercial. Uma sequência é um conjunto de uma ou mais visitas. Cada visita é composta por um *ID* de cliente, uma loja, o momento de detecção e outras informações relacionadas. Exploramos técnicas de *sequence mining* em diferentes representações. Com este estudo concluímos que estas técnicas podem ser usada para obtermos alguma informação contida nos dados. Encontramos evidências de que estas podem ser usadas para extrair informações em todas as representações testadas. Também discutimos outras informações e diretrizes que serão importantes para implementar esta solução numa base de dados de retalho.

Agradecimentos

Esta tese não seria possível sem a ajuda de várias pessoas.

Ao professor Carlos Soares, pela excelente orientação e todo o apoio que me forneceu e que me ajudou sempre em todas as dúvidas e novas etapas que iam surgindo ao longo do projeto, apesar do seu calendário apertado. Agradeço igualmente todos os fundamentos teóricos explicados, fulcrais para o sucesso desta dissertação.

Ao Zafeiris Kokkinogenis pelas incansáveis sessões de trocas de ideias para que eu conseguisse perceber melhor vários pontos fulcrais das temáticas estudadas. Também agradeço pelo interesse e confiança que depositaste em mim ao continuares a acompanhar o projeto.

À empresa Movvo, pela oportunidade, pelo tema da dissertação, pelo apoio que me foi fornecido e pela base de dados. Em especial à Diana Almeida e ao Paulo Castro pela disponibilidade demonstrada.

À minha namorada Susana, por me dar todo o apoio e carinho necessário. Obrigado também por me ajudares a encontrar motivação quando ela parecia não existir. Obrigado por seres quem és para mim.

A todos os meus amigos e colegas da faculdade, que sempre me apoiaram nos momentos mais complicados, em especial ao Rui Neves, ao José Taveira e ao Artur Ferreira.

Por fim, o meu muito obrigado à minha família por me possibilitar a frequência do curso e por me disponibilizar tudo o que tenho. Peço desculpa por todos os momentos em que fui muito chato e rabugento. Um agradecimento eterno e especial à minha avó e à minha mãe que me aturam e me ajudam em tudo o que preciso. Um agradecimento especial também ao meu pai que, apesar de discordar várias vezes comigo, me compreende sempre e ajuda-me a ser uma pessoa melhor. Obrigado a todos os outros membros da família que sempre me apoiaram e disponibilizaram-se para me ajudar no que fosse preciso durante a dissertação e sempre.

"Every second of every day, our senses bring in way too much data than we can possibly process in our brains."

Peter Diamandis

Contents

List of Figures

Abbreviations

GUI	Graphical User Interface
RFID	Radio Frequency Identification
SPAM	Sequential Pattern Minning Algorithm
SPMF	Open-source data mining library

Chapter 1

Introduction

We live in a world where there is an increasing influence of technology in our lives. Every day we come across with new technologies and devices that, in one way or another, influence our way of living. Because of this, there has been a steady increase in available information and in the use of technologies. The evolution of storage technology also follows this pace and it is becoming easier to store large amounts of data. For example, in 1985 the cost of a GigaByte was in the order of 500 000 USD and it is currently of approximately 0.10 USD [MT03]. With the constant evolution of technology and the lower costs of storing information, it is evident that we are witnessing a boom of information in the XXI century.

The access to all sorts of information it is getting easier so people and companies are trying to acquire the maximum possible amount from it. Given the quantity and diversity of available data, it is not always easy to extract valuable information from it. Knowledge extraction techniques are used to find important hidden patterns and exceptions in data. Those techniques are often called data mining techniques. One specific type of data consists of sequences of events (e.g. clicks on websites). The analysis of this type of data requires specific methods, referred to as sequence mining techniques.

One domain where the growth in volume and diversity of data has been particularly significant is the retail industry. Among other problems, store and product differentiation have always had a great impact in this business. Studies show that the quality of the stores and the service provided influence individuals to buy more [CdL03]. One way of differentiating is the constant use and adaptation of new technologies in the business. The use of new technological developments in physical stores can potentially improve the customer's relationship with the brand and boost its sales. The gap between technology and physical stores has been bridged over time through the development of projects such as applications of on-line purchase of products, analysis of the impact of promotions, the creation of newsletters, and many other examples. Those approaches led to an increase of purchases and an increase of quality in the buying experience of e-commerce retail.

However, facilitating the purchases of consumers is not the only effective way to use the technology available. In order to get more people to buy their products, companies should also understand as well as possible their customers and their behaviour. The use of new technologies can also lead to a better understanding of the clients. If retailers know the profiles of their customers, that will allow them to better sell their products as well as satisfy the customers.

This project was carried out in collaboration with Movvo[1]. This company is specialised in detecting buyers behaviour and providing important information to retailers, shopping centres and marketing professionals. Through an innovative system of indoor detection of people, the company can collect Spatio-temporal data of people moving in places where the system is deployed. Thus, it has access to a large amount of data related to buyer location in a physical space and that can be very useful for the owners of those locations to gain a competitive advantage over their competitors. In order to extract the best possible information contained in this data, the company uses various data mining techniques.

1.1 Motivation and goals

The main goal of this project is to explore the use of sequence mining techniques for the analysis of data about customer visits in a shopping centre, as collected by Movvo's system. This analysis can potentially add important knowledge to the owners of shopping centres or specific retail stores. It also may provide knowledge about the trajectories that customers follow inside a shopping centre, founding the most popular combination of stores. This can be useful for several reasons. The location of a store is many times referred to the most important decision in terms of future success of the business [KCK02]. If we analyse some of the most frequent subsequences of stores we can understand the relation that a store has with the others in the clients' visits. For instance, knowing that a lot of people are attending two different shops, could lead to the repositioning of those two stores, with the goal of increasing the flow of people in certain areas. This could lead to increase the possibility of more purchases in the stores of those areas.

The ideal space where future stores in the shopping centres will be placed can also be a possible object of study by the analysis of this data. With a previous sequence mining analysis in other shopping centres, we could put new stores in an area where there are stores with a stronger relation between them to increase the convenience of buyers. We could also place it at some distance from them in order to increase the flow of customers in the intermediate stores. After that analysis, we could choose the convenience over the flow of people or vice versa. The location of future stores and the influence in the flow of people in a shopping centre are only two examples of the possible use of the sequence mining techniques. The potential for the use of this analysis in other situations and contexts is huge.

[1] https://movvo.com/

The information contained in retail-related datasets is not restricted to their location. Many other factors are present in them like the time they happened, for example. These datasets can be used to build a richer representation of the sequences. With these different representations, more interesting analyses can be performed over the datasets, leading to results in a different spectrum.

A simple example of the type of knowledge is the following. If we know that a considerable percentage of buyers in a shopping centre who go first to brand *X* and then to brand *Y* go after that to brand *Z*, this may be interesting both for the owners of these stores and to the owner of the shopping centre. The owners of the stores can have a better understanding of the behaviour of their own buyers and make deals with the other related stores to boost the sales even more. The shopping centre owner can have a bigger picture of the clients and take advantage of selling hotter store's sequences locations with better prices, for example. We can say that having access to this type of information is potentially quite advantageous. The way the owners of commercial spaces and shopping centres may or may not use that information to gain a competitive advantage over their competitors is out of the scope of this project but the inherent value of this information can be valuable.

1.2 Project

In this project, we analysed data provided by Movvo, containing sequences of customer visits in a Portuguese shopping centre where the client tracking system is already implemented. Each one of these sequences represents the path of a visit to the shopping centre of one of the buyers. Each element in the sequence represents a single place that was visited and the sequence is sorted in temporal order. For example, if a buyer enters in the shopping centre and visits brand *X* first, then brand *Y*, then brand *Z* and finally exits the shopping centre this sequence of events will result in a data sequence similar to: **< Entry - X - Y - Z - Exit >** We applied sequence mining algorithms to analyse these sequences with varying degrees of confidence to look for frequent subsequences. The sequences represent common paths followed by the buyers. That information, along with other processes, were used to find buyers behaviour in the shopping centre.

This project was divided in experiences with different representations. Each representation contained different elements. This diversity of elements in each experience led to a complete analysis of the retail data. By dividing the experiences in representations we were able to detect different types of information related to each type of elements.

There are several techniques for sequence mining. Each of these techniques has the same goal but are quite different both in its implementation and in the input format of data and output format of patterns. Different techniques were considered and the ones that fit the problem best, namely concerning the different possible types of representation of the sequence, were used.

Before the application of the sequence analysis algorithms, a significant amount of effort was dedicated to preparing the data. The input required for the execution of each sequence mining technique is often different and so it was important to properly format the data for the selected algorithms. In addition, the data provided was in a raw format and, therefore, had to be refined. For instance, it is necessary to remove information which is incomplete or is not suitable for a given technique.

The definition of the appropriate threshold support is the most important factor on the quality of the results. It was also very important to define the minimum size of the sequences and the minimum number of stores that will be considered in a visit. This limitation of visits was used to refine the results.

The implementation of the sequence mining algorithms led to the discovery of various patterns. This information can later lead to potentially obtain a better understanding of clients' habits and to the discovery of some of the buyers' intentions in a shopping centre. When you are doing sequence mining analysis in clients' visits there are some patterns that are quite expected to be found like a sequential visit of various shops of the same type (various female shoe stores in sequence, for example) or a sequential visit to stores that complement another purchase (going to baby clothing store and after that to a toy store, for example). In some cases were found some patterns that seemed strange at first sight. This kind of patterns provided important new information in the study of the sequence mining analysis and in some case they proved to be an important eye-opener to new perspectives.

Chapter 2

Basic Concepts and Related Work

In this chapter, we discuss the basic concepts of the areas that are related to the work described in the dissertation and we present some of the main directions of research in sequence mining and behaviour prediction. Although not being applied in the dissertation, behaviour prediction is discussed as a possible element in a future improvement of the analysis. The technological and scientific choices that we made were based on this discussion. Furthermore, we identify some opportunities that this work presents.

2.1 Data Mining

Every day, about 2.5 quintillions (2.5 x 10^{18}) bytes of information is created in the world and every two years the amount of stored information doubles [Big12]. The vast majority of the current databases are large and can contain much information. Analysing by hand each element to look for valuable information would be unaffordable. Due to this huge growth of stored data, some software and techniques are not capable of managing and processing it in a reasonable amount of time.

Data mining consists of a set of techniques that will extract non-trivial, previously unknown, but useful hidden information from large databases and in the fastest time possible [HPK06]. Mayer-Schonberger and Cukier also states that data mining is the term used to refer to things that we can do on a large scale, that are not possible to do on a smaller one, by extracting new information or create new forms of value [MSC14]. In order to improve and organise the processes related to data mining, Labrindis and Jagadish [LJ12] propose a process model represented in figure 2.1 divided into two main phases: data management and analytics.

The data management phase includes the processes that are leading directly to the modification of data. These processes are concerned with gathering and storing the data, refinement, extraction of information, integration, modelling and retrieval. The analytic phase is where the extraction of knowledge from the data occurs. This phase includes all the representations of the analysed data

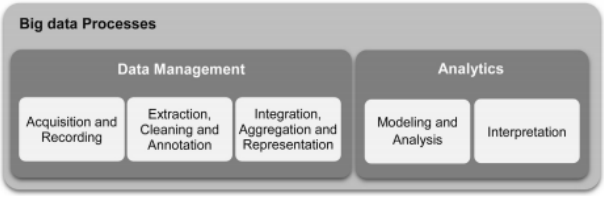

Figure 2.1: Big data main processes [LJ12]

and the conclusions taken from it.

Data is available in diverse formats: text, audio, video, social networking data, etc [GH14]. Even when the format is the same, there are a lot of aspects that distinguish two different databases. For example, in a database the different attributes of each element can be separated by point, semicolon or tabs. To process different databases, we need different approaches. There are techniques and technologies that are specific to different types of data. Because of that, the selection of tools and technologies to use in a data mining project will be essential for the success of the project.

A good example of a possible use of data mining could be described in a search engine: Google, for example. Every day, people submit in this platform hundreds of million of queries. If we think each one of those queries describes the information a user needs, we can gather that set of queries in a huge collection and detect behaviours from people that can lead to some conclusions. For example, in 2016, Great Britain made a *referendum* about the country membership in the European Union. The day after the win of the "Leave" vote, due to the use of data mining techniques on the queries, Google claimed that the first five trending searches in the United Kingdom were about the European union [Bre16]. That fact reveals that there were a lot of people in the United Kingdom that did not search those queries before the election, but only after the results. This could lead to several questions as "Were people well informed enough for voting?" or "people did not really think that leave option could win". Other application of data mining techniques in Google could also be the prediction of a disease. If the frequency of the search of queries related to the flu increases substantially, it is expected that a lot of people are sick. This way, Google can predict up to two weeks before the occurrence of a flu epidemic than the medical institutions can [HPK06].

Youtube, the largest platform for sharing videos in the world, deals with thousands of video uploads every minute. Storing video data is more complicated than other types of data because they occupy more space. This size could potentially lead to a more complicated data mining because the larger the size of the data, the greater will be the time of processing it. Nevertheless, there are implemented recommendation systems, lists of the most trending videos at the moment, etc. These features that are present on Youtube, come from the use of specific data mining techniques that try to select the most relevant results for each user [CDL08].

Data mining techniques can be virtually applied to every dataset and its use can be very helpful

to all sectors of a company. Factors like the time available, the domain of the data set and the selection of desired fields of information must be very well controlled because a bad consideration of them can lead to wrong results. It is very important to study different factors of the dataset that you want to apply the data mining techniques before you start applying them.

2.2 Data Mining for Retail

Retail is a fundamental economic activity. All of the stores that sell a product or a service, big or small, are considered to be in the retail business. This sector of the economy is normally very competitive. Usually, every store or retail company tries to differentiate themselves from other stores by changing product prices, increasing the convenience of the clients, offering promotions, etc. Thus, retailers must know very well their customers and their habits to be able to make the right decisions.

The increase of the population living in urban areas has led to an increase in the number of shopping centres in the world. This kind of commercial spaces can be very convenient to people, gathering several types of stores in one place. But, due to the fact that there are many shopping centres that compete with each other and all of them want to be the best, they need to find ways to gain a competitive advantage over the others.

Using data mining in shopping centres can be very beneficial to them because they will be able to extract new information about their business and about the stores. There are several processes in retail that could potentially benefit from the use of these techniques and that are already being used. For example, the location of a store can be a very important factor of success and the changing of it can be, most of the times, irreversible [CTK13]. If we could predict an ideal place for a store we could make the store generate more money and increase their chances of becoming successful. Theoretically, if a store is in a place that has a good access and it is near the store's target audience it will sell more. By applying data mining techniques to the data from previous stores, the owners can understand what are the most important factors for the success of the business and then choose the best location.

Retail has undergone a lot of changes due to the introduction of new technologies in society. Retailers have to adapt to these changes in order to keep up the pace with the technology development and with the new companies selling online and not physically. Knowing that two thirds of mobile phone users use or have used mobile phones to buy a product online [CCS14], one of their main effort is to create mobile and web applications that could compete with already existing online stores. The retailers are continuously adapting their processes in the physic stores and creating online marketing campaigns to adapt to the growing use of new technologies. But even so, there are quite a few ways to improve their adaptation. One of these ways is to apply different

data mining techniques to different types of information.

If a retailer knows where her customer have been in their store she can learn important information about their profiles and habits. The access to that information can lead to a better understanding of those customers and an increase in the profit of the retailer. This kind of solution is not only being used in retail but also in tourism, mobility research and many other commercial areas [XCA05, Lie13]. There are also projects that already have the ability to collect customer data in the stores in order to profile them with patterns and clusters with attributes as age, gender, locations visited inside a shopping centre, etc. Lin writes about a system that is able to detect people using WIFI-based indoor localisation technology [Lin13]. This technology is able to detect the movement of people in a shopping centre. After one week of data gathering, the author used data mining techniques to implement a recommendation system allowing the owners to better predict where their customers would go next.

There are a lot of opportunities and possible ways to improve the use of data mining in retail. Despite being already in use in many places, there are many possible future applications for them in this type of business. The retail stores that will invest in them in a short term may benefit with its use and gain a competitive advantage over their competition. This investment could be made in improving the current features or developing new ones.

2.3 Sequence Mining

Each data mining method is able to identify a restricted type of patterns. For example, With association rules [Agr], we can formulate a rule X => Y, meaning that when X occurs it is likely that Y also occurs. With its use it is not possible to obtain patterns such as Buy A implies Buy B within a certain time interval or people Buy C every week [ZB03]. However, a lot of data includes this type of temporal information that is interesting. Often these temporal relations represent sequences that are interesting. These sequences are a set of ordered events that are recorded in a database. The goal of sequence mining techniques is to analyse a large set of data with a sequence format and extract the maximum possible similarities between them. [MSW14]

2.3.1 Definition

Generally speaking, we can say that sequence mining is the use of a method that has the goal of finding patterns in a database that contains data in sequence format [CTG12]. The minimum similarity level between the candidates and the original sequences that we are analysing when we use sequence mining is called support threshold. To set this value, the user needs to find the level that he considers satisfactory for the algorithm to find only patterns with a similarity level above it. For example, if we define a threshold of 10%, we will receive all the sequences that were present in at least 10% of the original sequences in a dataset. Sequence mining is often referred to as

sequential pattern mining.[ZB03]

Time is a very important and decisive factor in sequence mining. The elements of a sequence have to be organized in order of occurrence. So it is mandatory to order the items that are contained in the sequences. Time can also be a factor of further analysis if we include it as a factor when detecting sequence similarities.

A sequence can be defined as an ordered list of itemsets a_i denoted by $< a_1, a_2, \ldots a_n >$, with n $\subset \mathbb{N}$ being the last member of the sequence . A sequence with k itemsets is called k-sequence. An itemset can be defined as a set of items s_j, $\{s_1, s_2, \ldots s_n\}$. Each item can only occur one time in each itemset, but can occur multiple times in a sequence. A sequence S1 $< a_1, a_2, \ldots a_n >$ is contained in a sequence S2 $< b_1, b_2, \ldots b_n >$ where integers $i_1 < i_2 < \cdots < i_n$ exist such that $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, ... ,$a_n \subseteq b_{in}$. For example, sequence < {a} , {b} , {c,d} > is contained in the sequence < {y} , {a, f} , {b} , {e, f} , {c, d, x} >, whereas < {a , b} > is not contained in the sequence < {a}, {b} >. The first example implies that a and b are concurrent while the second implies that b occurs after a [Oli15].

2.3.2 Sequence representations

The information contained in a sequence can be very diverse. Sequences' structure is very modular and this allows us to choose different ways to represent the same information. So it is very important to correctly define the content that we want to add in order to better decide a fitting sequence structure to use later. This fact happens in the sequences because each sequence is a set of itemsets by itself. For a better understanding and consistency, a sequence S will be represented as < {a} , {b} , {c} > where the elements {a}, {b} and {c} are itemsets.

For example, if we want to register the people that passed by a certain location over time we can use a sequence to represent it. Let us imagine that 3 people P1, P2 and P3 passed in that location by this temporal order. We can represent this sequence with < {P1} , {P2} , {P3} >. Each members of the sequence is an itemset so we could store more information in each one. For instance, we can easily add information such as age or the weight of the person in each itemset. The first person is 23 years old and weighs 60 kilograms, the second is 36 and weighs 72 kilograms, and the third is 52 years old and weighs 82 kilograms In this case the sequence could be something like < {23, 60} , {36, 72} , {52, 82} >. This way the data in sequences can become very detailed and contain different sources of data about each element.

The elements belonging to each itemset can also have different formats. Each one of them will keep the other attributes and add the new one. For example, we can add a string for each person representing the place of residence. If we continue with the former example the new sequence could be: < {23, 60, London} , {36, 72, Liverpool} , {52, 82, Manchester} >.

2.3.3 Basic Algorithm and SPAM algorithm

All the sequence mining algorithms have a similar structure. The first step is to find all the possible sequence candidates. Different sequence mining algorithms use different ways of choosing valid candidates according to the criteria chosen by the user. In the second step, we calculate the support of the found sequences. All the candidates that are at least equal to the minimum support threshold are identified and gathered in the returned solution.

Figure 2.2 presents the pseudo code of one of the first set of sequence mining algorithms: Apriori-Based algorithms. The structure is very similar to the most recent algorithms. First, the algorithm stores all the possible candidates of being sequences in L_1 (step 1). A subsequent iteration k is divided into two phases. In the first one, the large itemsets L_{k-1} found in the $(k-1)$th iteration are used to generate the candidate itemsets C_k (step 3). In this step, an apriorigen function is used to select all the possible candidates. The main goal is to prune the candidates that not fit in a predetermined criteria. After this, the database is scanned and the support of the selected candidates C_k is calculated(step 7). This selection results in L_k that is the subset of C_k but only with candidates that have an equal or higher frequency than the minimum support threshold.(step 9) The sequences found by the algorithm will be the union of the values contained in L_k [AS94] (step 11).

The essential step for the algorithm to be more efficient is the apriorigen function. For fast counting, we need to efficiently determine the candidates in C_k. This is the step that keeps evolving since the creation of the first sequence mining algorithm [AS94]. As new techniques have been applied to other algorithms types of sequence mining, the improving of this function as been the main difference between them.

There are two general properties that are applied in the selection of the candidates in the sequence mining algorithms:

- if the sequence < a, b, c > is frequent, so is < a, b > and < b, c >
- if a sequence S is not frequent, then none of the super-sequences of S is frequent. Ex: If < a, b > is infrequent, than < a, b, c > is also infrequent.

1) $L_1 = \{\text{large 1-itemsets}\}$;
2) **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
3) $C_k = \text{apriori-gen}(L_{k-1})$; // New candidates
4) **forall** transactions $t \in \mathcal{D}$ **do begin**
5) $C_t = \text{subset}(C_k, t)$; // Candidates contained in t
6) **forall** candidates $c \in C_t$ **do**
7) $c.\text{count}$++;
8) **end**
9) $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
10) **end**
11) Answer $= \bigcup_k L_k$;

Figure 2.2: Pseudo Code of Apriori algorithm [AS94]

SPAM algorithm is one of the sequence mining algorithms. It is considered one of the efficient algorithms for mining sequences [AGYF02]. This algorithm has a generate-and-test feature to test all the possible candidates. With this algorithm the memory consumption is reduced by an order of magnitude related to the size of the database.

2.3.4 Algorithms Types

In terms of sequence mining, we can refer to three main groups of algorithms: Apriori-based, Pattern Growth and Early-Pruning [ME10]. Each one of these groups contains several algorithms that try to reach the same solution but through different approaches.

The Apriori-based algorithms use a *generate-and-test* type of approach. They select candidates and then test them. They are based on the premise that if any sequence X is not frequent, then another sequence that contains X will not be frequent, and thus, it will be possible to enhance our choice, decreasing the number of potential solutions. This group of algorithms were the first to emerge.

On the other hand, Pattern-Growth-based algorithms use the generate incremental approach. During the incrementation, these algorithms use methods of divide-and-conquer that will make more accurate projections in order to reduce the number of possible sequence candidates comparing to the previous ones. This type of algorithms arose from the need to find different approaches of Apriori-based algorithms that were often rudimentary and not very optimised.

Finally, in most recent years, approaches that attempt to prune candidate sequences as soon as possible are being developed, thereby making the process more efficient. These algorithms are called *Early-Pruning* algorithms and they are very similar to the Growth Pattern-algorithms, with the exception of the sequences candidate selection process. In these algorithms are applied more recent techniques that allow these algorithms to be faster in this process comparing to other sequence mining techniques.

Figure 2.3 presents a complete taxonomy of the existing algorithms and methods of sequence mining. In this organisational chart, we can see the algorithms divided into three main groups: Apriori-based, Pattern Growth and Early-Pruning. In each one of the groups, we can also notice a division of the algorithms based on the approach used and theoretical basis. Some of the algorithms belong to more than one of the branches. For example, the SPADE algorithm uses a temporal-vertical database as a temporary memory and it uses a lattice transversal method as a process of visiting each node in a tree data structure.

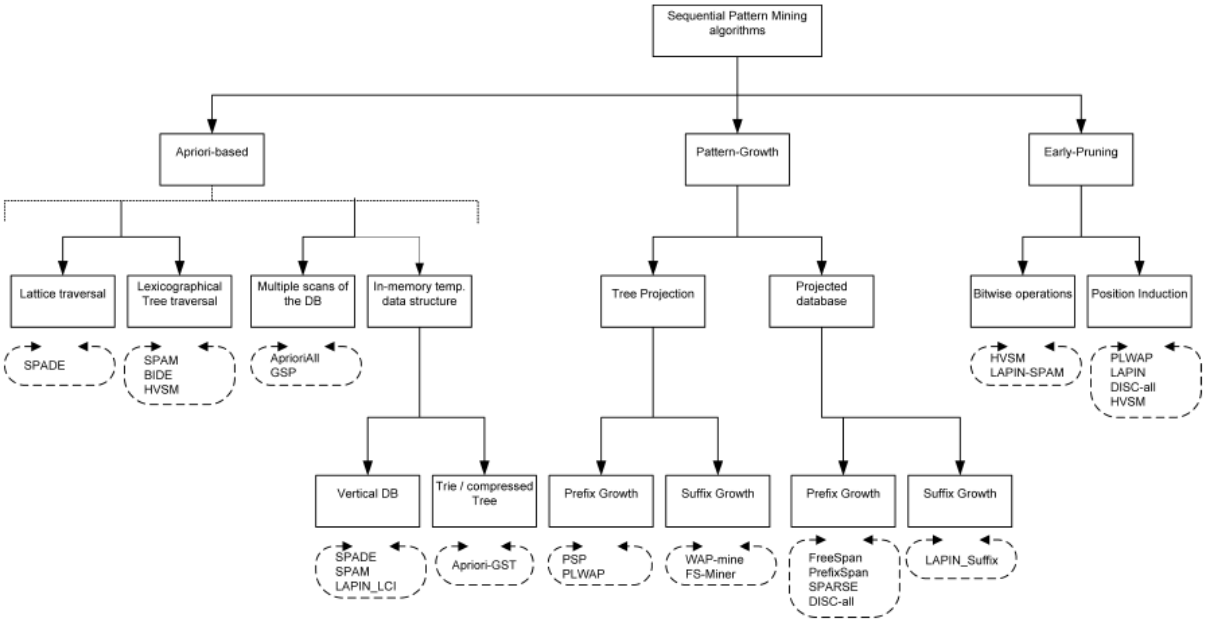

Figure 2.3: A taxonomy of sequence mining algorithms defined by Mabroukeh and Ezeife. Reproduced from [ME10]

2.4 Related Work

Many projects used sequence mining techniques in recent years in different domains. In this section, we will mostly discuss projects that used sequence mining and are somehow related to retail and sales or to a similar area of interest.

Cabanes et al. propose a system of detecting human path behaviour inside a physical retail store with a powerful tool that detected homogeneous sub-sequences of spatio-temporal data collected by a RFID system [CBDJ09]. Cabanes et al.'s project used sequences related to the locations of the clients. After the sequence mining analysis on the data, the authors propose a system that helped define different preferential areas of the clients inside a physical store. With those areas he did a study on the behaviours in the most frequently paths.

In 2006, Eichinger et al. used sequence mining techniques to find customer behaviour predictions on the field of telecommunications [ENK06]. The dataset was composed of real customer data from a major European telecommunication provider. Each sequence represented a time ordered set of actions that one client made. For example, one sequence could be something like < {call, 9126} , {sms, 2543} , {video-call, 3222 }, {call, 3222 } >. In this project the authors combined sequence mining techniques with decision trees to build a classifier for each sequence of data in order to divide the data in branches and enable the analysis of separate segments of clients. This also allowed the division of sequences with single events that needed to be separated from the others. These sequences had to be separated in order to not negatively influence the results of the other sequences that contained more than one element. The authors concluded that the classic definition of support threshold and usual sequence mining techniques were not enough to reach customer behaviour prediction. However, they were capable of finding potential relationships that could later lead to that goal. The application of sequence mining in the telecom industry has also been studied by people from Samsung Research America [MSW14]. The main goal of this project was to make the phone understand frequent sequence patterns from the user. Each sequence was composed with actions performed by the user in a mobile operation system. The project resulted in a new functionality that gave users the possibility of enabling applications that were chosen more often in a similar usage context, improving the interaction with the phone and the user experience. The developed system used was the Mobile Sequence Miner (MSM), an algorithm adapted to this context but based in the PrefixSpan algorithm.

Another project was the one developed by Bermingham and Lee [BL14]. Their goal was to extract spatial-temporal patterns from data from Flickr, a social network of photo sharing. Their work consisted in the detection of the most frequent sequences of visits by tourists to different regions in Queensland, an Australian state. They used a framework called sequential pattern mining framework to analyse the data. The use of this technique help the researchers to conclude where and when people were going and where they were likely to go next. The division in regions-of-interest and the consideration of space and time simultaneously made this work particularly interesting. The use of sequential pattern mining algorithm is also verified in Tsai and Lai's work [TL15]. In this project, the authors tried to identify spatial-temporal behaviours from the visitors of a simplified theme park. They used a type of sequence called Location-Item-Time sequence that uses locations from the park. They developed an algorithm that had the objective of finding the most popular locations in sequence called Location-Item-Time PrefixSpan algorithm, as an

adaptation of the PrefixSpan algorithm. The use of this algorithm help to prove the possibility of developing a system that could help managers to better understand visitors experiences and eventually discover behaviour intentions.

A probabilistic database framework was also developed by Muzammal [Muz12]. The objective of this framework was to find the most certain option in situations of the uncertainty of a source associated with an event. They first used dynamic programming to compute the probabilities and then use a breadth-first algorithm (similar to GSP) and depth-first (similar to SPAM) algorithms that generated the best candidates. These GSP and the SPAM algorithms are included in the figure 2.3 and are both apriori-based. The results showed optimisations in CPU cost compared to other conventional methods.

Sequence mining was also used in some projects in the pharmaceutical industry. Wright proposed the use of sequential pattern mining to predict the next medication to prescribe [WWMS15]. The CSPADE algorithm was used to mine sequential patterns of diabetes prescriptions. The study concluded that CSPADE algorithm was effective in predicting the next steps in a patient's medication regimen because the technique is an efficient tool to discover temporal relations between different medications.

Goel and Malick developed a system of detection of customer purchasing behaviour using sequential pattern mining [GM15a]. The main goal was to understand the behaviour of bank clients. It was expected to help to improve the efficiency of the bank in terms of creation and management of accounts. In this project the PrefixSpan algorithm was used, with the FP Growth algorithm to find candidates. The K-means algorithm was used before the application of PrefixSpan to divide the clients into clusters.

2.5 Behaviour Prediction

Behaviour Prediction is the study of predicting future events that will happen within a determined scope. This prediction can be obtained with different activities. The use of analysis of data and machine learning activities are two of them [Abb15]. This kind of information is not easy to obtain and it is difficult to be somehow accurate about the conclusions. This happens when the conclusions are only based in one activity. If we want to be more accurate we need to add several kinds of data during the behaviour prediction analysis. For example, if we want to analyse the behaviour of the clients that receive a call from a call center that sells a product, we can perform data mining in their interactions and evaluate the success rate. But only this data can not be enough to predict behaviours from their future clients. For example, if we have a tracked register of the conversations we can analyse in what point of the conversation the people rejected the product. Knowing this kind of behaviour the call center company can try to improve the process in those

specific parts of the conversations to make a better persuasion job in the future calls.

There are a lot of areas where behaviour prediction already happens such as psychology, economics or marketing. Adapting common behaviour prediction to human behaviour prediction can be very complex to do. In figure 2.4, we can observe an example of a model of human behaviour detection in a hospital. In this model, there are plenty of variants that can be discovered before we can reach to conclusions about the behaviour. There other systems of behaviour prediction that are more simple and straightforward. Because of that some of the factors that will determine the prediction of Human behaviour must be very well crumbled.

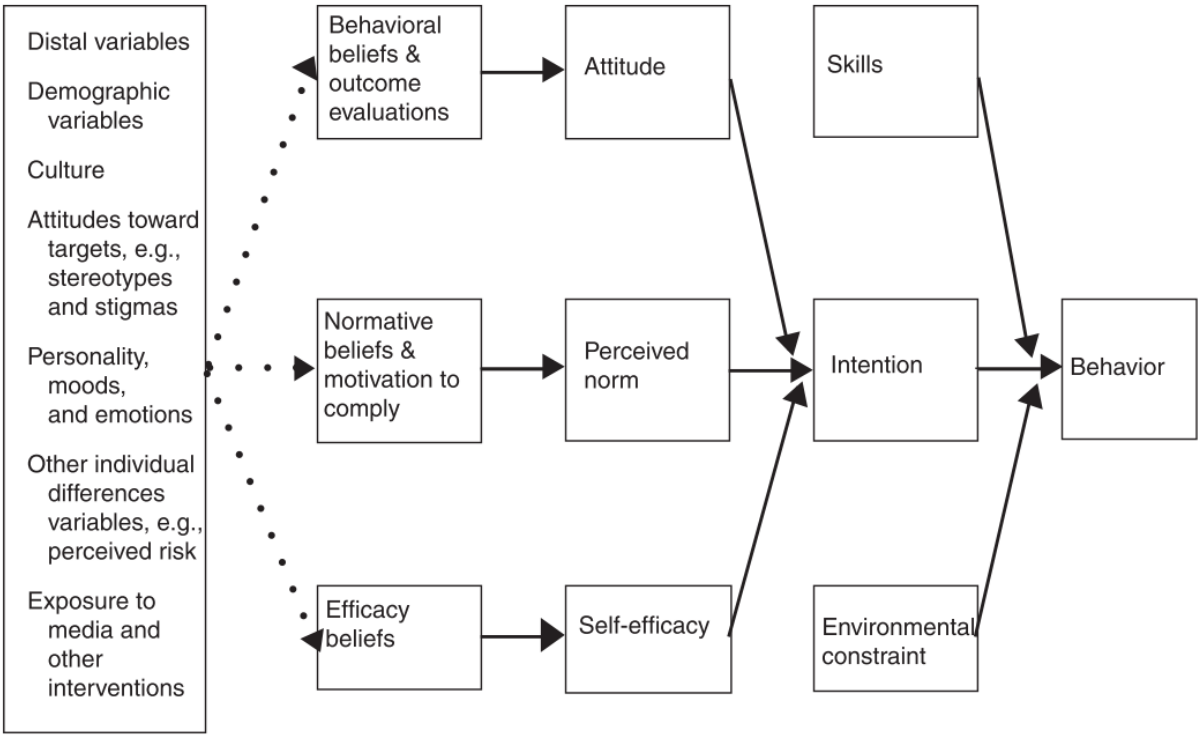

Figure 2.4: Example of a model of human behaviour prediction [FY03]

One important thing about behaviour prediction is the representation because, without a model that defines the process correctly, it is going to be very difficult to implement a successful project. One of the most often used graphical models is the Bayesian Network Model that represents the relationships between variables and condition dependencies in a system (Figure 2.5). The Bayesian Network Model is a type of model that represents several variables and the their conditional dependencies between them. There are also other people investigating new frameworks of representation to suit behaviour prediction better. This happens because there are usually multiple objectives and constraints in a project [Abb15]. One of those frameworks was proposed by Abassi and he claims that the implementation of this new framework could make retail stores to be ten percent more accurate and could lead to an increase of millions of dollars in sales in the e-commerce stores

analysed.

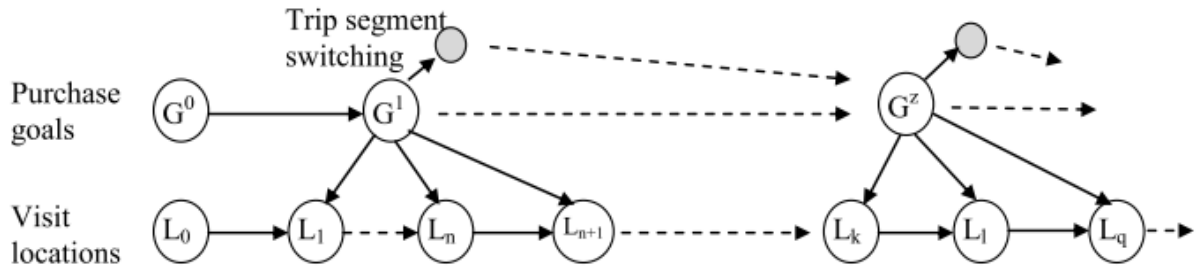

Figure 2.5: Example of a Bayesian model network modelling consumer purchases and travel choices [Yan10]

One example of a project that aimed the discovery of customer behaviour is the one proposed by Goel and Malick [GM15b]. In this paper, using some techniques of sequential pattern mining, the authors tried to find some purchasing patterns in a bank facility. With that analysis together with other data like the effective purchases of the clients, they concluded that it was possible to determine some customer purchasing behaviours. In this project it was made an analysis of sequence mining by comparing it with previous data provided by the bank that was not related to sequence mining analysis.

A very interesting sub-area in behaviour prediction is the nonrecurring behaviour analytics. It is motivated by the necessity to detect behaviours of a system that should be possible but for some reason, do not occur [CS15]. This may be due to different reasons such as bad system implementation or system hacking. It is interesting for areas with critical systems like banking, capital markets, government services and also consumer behaviour. One example of the possible advantages of the application of retail analysis is the detection of a boycott on some product or sales area.

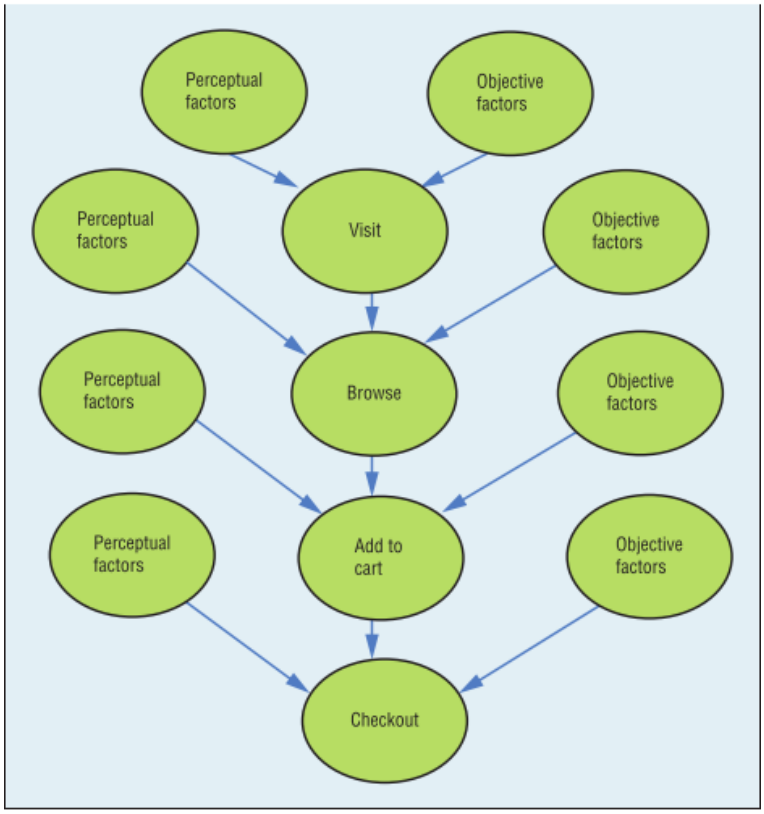

Figure 2.6: Example of a Bayesian model network for consumer behaviour prediction in e-commerce [Abb15]

Chapter 3

Case Study

In this chapter, we explore the different components executed in the project and justify the choices that we made. An analysis on the available data was performed to better understand it and to have the ability of comparing the results of the experiences with it. We also explain how the project was developed and the answers that could be answered with it.

Figure 3.1 summarises the approach followed in this project. The blue rectangles represent the most crucial components of the project and the ones we have implemented. All the project was repeated many times for each experience until we reached at least some satisfactory results. Each of the components was also improved during the duration of the project. The white rectangles represent the components that could later lead to a system that potentially could predict behaviour intentions of buyers automatically in the future. These components were not implemented due to the time restrictions.

3.1 Data description

All the data used in this project was provided by Movvo. Using data collected by antennas placed in shopping centres, the company manages to collect information about the movement of the people. The information is refined and then transformed to represent the visits of customers to the stores, at a given time. All store names and people are identified with a numerical identifier due to the confidentiality of the information.

The dataset is composed of 102110 records collected in a shopping centre in Porto. Each record represents one person passing by a determined place at a certain time. Each record contains not only the store but also the time, the duration of the visit and other parameters. These parameters allow us to add more extra knowledge to each record. With the time parameter, we can find the day of the week and the time of the day of that detection. With the location, we can uncover which store is being visited and its respective category. The field that represents a person is a unique ID that changes for each visit. If a person visits the shopping centre in another day

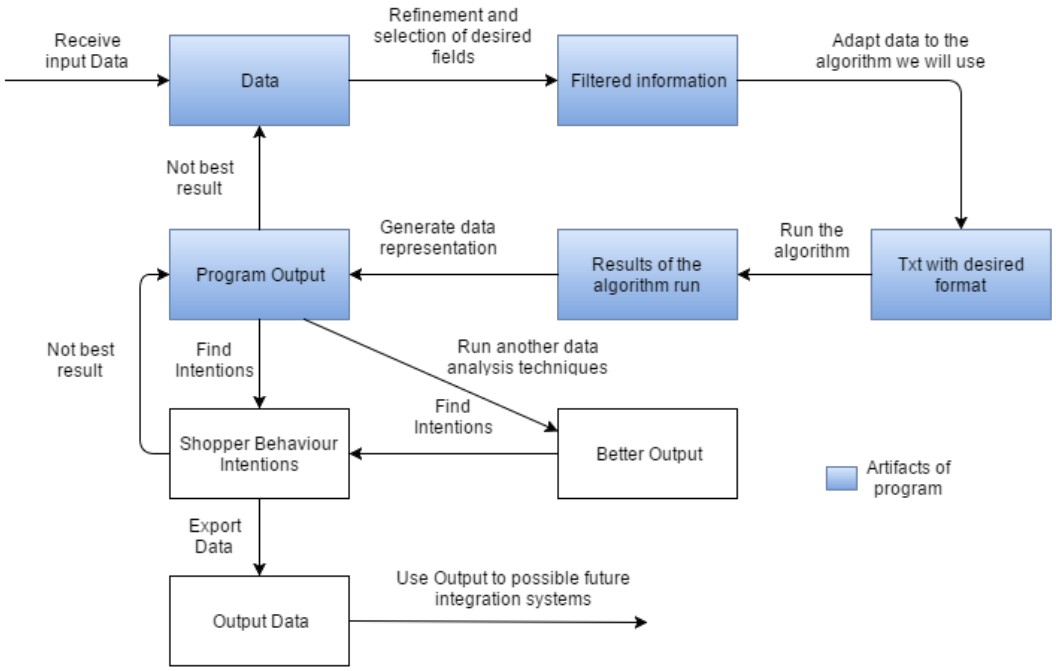

Figure 3.1: Diagram of the process that we have used during the dissertation. In blue rectangles are the most crucial phases of our project. In white rectangles are the phases that will not be implemented but could lead to an automatic system of sequence data analysis

this unique ID will not be held but changed for each new visit. The data also contains a field that represents the duration of a visit. For example, if one person was in Store A at 6 PM and left it at 7 PM we will have a record that describes that information and it will also contain a field with the duration 60 (minutes). In figure 3.2 we can see an example of four lines of data.

Date	Idclient	StoreId	Group	Subgroup	Activity	Dia	Duration
2015-11-23 10:03:02	f8a04a3da37bd7826a64f106485d809d027ac086	470	Miscellanious	Miscellanious	Miscellanious	Monday	1
2015-11-23 10:05:01	e2af8e9942cc105955847ed3d64351fa5c956ecf	361	FASHION, FOOTWEAR & ACCESSORIES	FASHION	CLOTHING IN GENERAL	Monday	1
2015-11-23 10:05:02	2fa5f74817c64232d8328cf5387f875d208a19fd	331	FASHION, FOOTWEAR & ACCESSORIES	FASHION	CLOTHING IN GENERAL	Monday	3
2015-11-23 10:05:03	2fa5f74817c64232d8328cf5387f875d208a19fd	332	FASHION, FOOTWEAR & ACCESSORIES	ACCESSORIES	JEWELERY & ACCESSORIES	Monday	2

Figure 3.2: Example of the information contained in the dataset

The data contain records of 180 stores that are divided into 48 categories. A visit of a person to the shopping centre is a set of stores visited in sequence by one person. After a preliminary analysis, we detected 45092 different client visits. This makes an average of 2.26 stores visited by each person per visit. The time interval is between 23/11/2015 and 29/11/2015. In figure 3.3 are presented the Top 5 stores of the data. These five stores were the stores that contained more records of visits in the analysed data. The store 536 stands out clearly from the others containing at least twice the visits of all the stores, except the stores 509 and 381. We can also observe to which category each of these stores belong. In figure 3.4 the Top 5 most frequent store categories are presented. Once again the range in the number of visits is very marked between the first and

the remaining categories. In the table we could also observe the relative frequency of each category. The figure 3.5 presents the distribution of store categories in each visit. This is a graphic that contained a much more detailed information of the number of visits to each category.

Code of Store	Type of Store	Number of visitors
536	CLOTHING IN GENERAL	5620
509	JEANS AND CASUAL WEAR	3233
381	BIJOUTERIE	2884
531	Miscellanious	2489
487	LADIESWEAR	2435

Figure 3.3: The top 5 Stores by number of visitors in the dataset

Type of Store	Number of visits	Percentage
CLOTHING IN GENERAL	17387	17,03%
MISCELLANEOUS	7305	7,15%
LADIESWEAR	6852	6,71%
JEWELERY & ACCESSORIES	6786	6,65%
BIJOUTERIE	6380	6,25%

Figure 3.4: The Top 5 store categories by number of visitors in the dataset

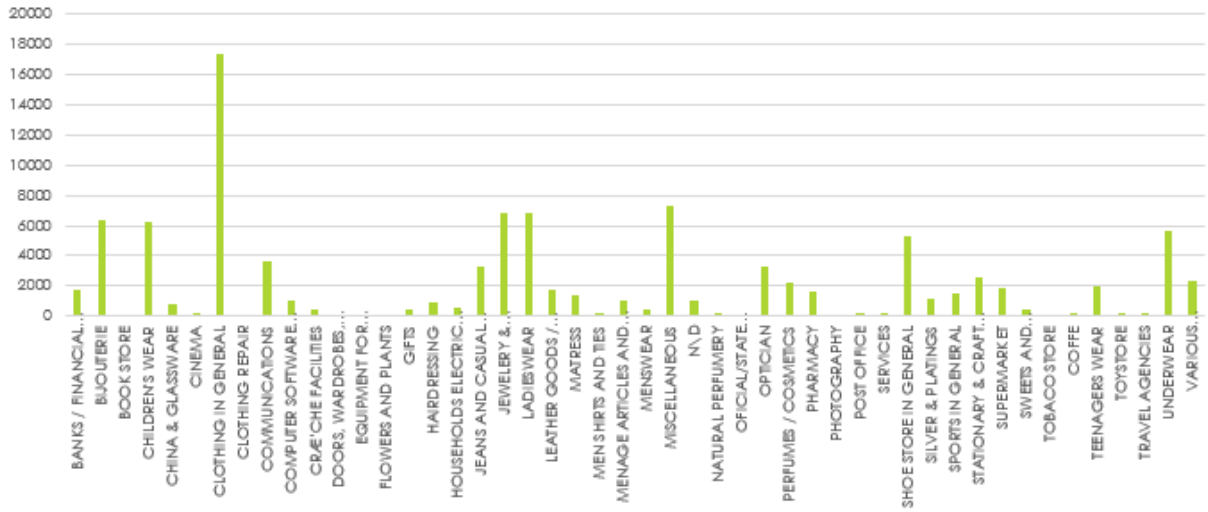

Figure 3.5: Distribution of the records by category

The dataset has some limitations. The first one is the amount of data. A little more than one hundred thousand registers may seem a very reasonable amount of data but in reality, it may not be enough to validate some more detailed results. The range of the records is also low, corresponding only to data related to one week, which is a short period. Some periods do not contain any data which further limits the extent of analysis. The records also show few visits to restaurants and film theatre combined (less than one hundred). Knowing that the shopping centre where the data was collected is composed of a large food court and several movie theatres, this is surprising.

3.2 Software and Algorithm

The proposed work has a high-impact practical component for the application context we study. Every test that we made was executed several times until we reach the desired results. A fast and efficient algorithm is an important requirement as the computational time has a great impact when working with large-scale datasets, as it is in our case study. The available databases contained a large volume of data that limit the number of experiments that could be carried out in the empirical study. For each scenario, several tests were conducted.

Various data mining software frameworks and algorithms are already implemented and available for free use. After an analysis of the existing tools and libraries, we have decided to use SPMF, an open source data mining library [FVLG+16]. The toolkit implements 122 pattern mining algorithms including sequence mining, which is the task addressed in this work. We choose it because it is very complete, with 26 sequence mining algorithms implementation, it is open source and contained a GUI.

We have considered several algorithms from the SPMF library that use the same input file format. The biggest input file contained more than 40000 lines. There were several algorithm candidates available that satisfied our pretensions and obtained similar performance results. Among the several candidate algorithms that fulfilled our analysis criterion, we chose the SPAM algorithm because because of the flexibility it provides. SPAM was the first implemented depth-first search strategy for mining sequential patterns [AGYF02]. With this SPAM implementation, we can choose the minimum and the maximum pattern length. The minimum and the maximum pattern length can help to limit the number of undesired patterns returned in the analysis. For example, we had not interest in analysing patterns of the results with only one element. The use of these variables during the sequence mining can lead to show more correct results because it widens the spectrum of search.

3.3 Research Questions

The main objective of this work was to analyse the sequence mining techniques applied in the shopping data provided by Movvo. The purpose of this analysis was to discover new informations

about the sequential visits of the buyers. It was also to confirm information about the buyers that could be obtained with other analysis.

With a sequence mining approach, we are able to find the patterns that occur sufficiently often in the data to be of any practical interest. Those patterns enable us to better understand the clients of a retail establishment. For example, if we know that there is a sequence A - B - C with 5 % frequency we know that a significant proportion of our clients visit those stores in that order. That information can lead to the discovery of, for example, popular routes of stores in the clients visits to that shopping centre. This analysis can also relate stores that are often present simultaneously in different patterns. It can also be used to confirm the general idea of the popularity and importance of those stores to the retail establishment that can be obtained using association rules. Through the analysis of the patterns, we can also speculate: is store "C" the place where the shopper frequently goes first when they want a specific product? Are stores "A", "B" and "C" of the same category and the client only purchases a determined product in the last store? These speculations, together with other data can lead to the discovery of important information about those stores.

The flow of people who go through frequent shopping paths can also be analysed through the patterns. Knowing that two stores are present in various patterns tells us that there is likely a strong connection between these two stores. Understanding this connection may lead to business value. For instance, store locations may be changed. If the goal is to attract more people to a particular area, we can distance those two stores, forcing people to go through areas between them. Alternatively, these two stores may be placed closer together to increase that connection, as well as customer convenience. This change can be made either in existing retail establishments or even in the planing of new ones.

The stores that appear more often at the beginning and at the end of patterns is one of the explored subjects in the dissertation . This positioning at the extremes of patterns may raise some questions. The stores that appear first may be the most attractive or they are not able to satisfy buyer needs, thus causing them to look into other stores. On the other hand, stores that appear most often at the end of patterns may be the ones that helped buyers to meet their needs or they may be less interesting. These questions can be further investigated, when detected by the sequence mining algorithms.

3.4 Representations

As discussed earlier, the data available contained information that can be used to build richer sequences, not only sequences of the stores that were visited by customers. One of our goals was to explore if the algorithms available were able to explore that information and extract patterns

that provided alternative and complementary perspectives of sequences. We have divided the experiences made in four different representations. Each representation is composed with sequences that contain different information. This division allowed us to analyse different aspects of the available data and reach conclusions on each of these representations. The first representation was composed of stores. The second representation was composed of the stores and the duration of the visit in that store. The third representation was composed of the stores and the time of the day of the visit. The fourth representation was composed with stores, the duration of the visit in that store and the time of the day of the visit.

Given that the implementation of the SPAM algorithm that was used is only able to deal with numerical data, we mapped the values of nominal variables into integers. The 100 stores contained in the data are represented with the numbers between 100 and 1175. This encoding was already in the original raw data. Based on expert knowledge, the duration of a visit to a store was divided into four categories:

- **Very short visit** - $visit < 5$ minutes: represented as 0
- **short visit** - $5 \leq visit \leq 10$ minutes: represented as 1
- **medium visit** - $10 < visit \leq 25$ minutes: represented as 2
- **long visit** - $visit \geq 25$ minutes: represented as 3

The *very short visit* represent visits that are too short to enable any useful activity in the store.

The time of the day that a visit was done is also divided into categories:

- **morning period** - between 10 AM and 1 PM - represented as 10
- **early afternoon period** - between 1 PM and 4 PM - represented as 11
- **afternoon period** - between 4 PM and 7 PM - represented as 12
- **evening period** - between 7 PM and 10 PM - represented as 13
- **night period** - between 10 PM until closing hour - represented as 14

This division was also based on experts knowledge and in the opening hours of the shopping centre.

In addition to the parsed variables being encoded it is also necessary to add tool-specific symbols that will be used to run the sequence mining tests. The separation between items of a sequence is encoded as '-1'. The end of a sequence is encoded as '-2'.

The first representation is the simplest because it s the only of the analysed representations that only contain one value in each item of the sequence. These sequences represent the stores

that someone visited in temporal sequence during a visit to this shopping centre (figure 3.6). Each store is represented by a predetermined code. This representation allows an analysis of the stores that are more common visited in sequence. It also allows the discovery of stores that are often related in several sequences.

```
366 -1 478 -1 495 -1 502 -1 536 -1 -2
310 -1 311 -1 529 -1 -2
323 -1 380 -1 396 -1 458 -1 -2
```

Figure 3.6: Representation of stores' sequences

Each item of a sequence can be represented with multiple values. In this project, each item represents a visit to a store and the characteristics of those visits will be included in the representation. In figure 3.7 are represented a few examples of these sequences. Besides the ID of the store, each item also contains the duration that the client spends in that store. For example, in the first line of figure 3.6 the second item contains '1 509', which means that the customer visited store '509' and the visit was a very short one, represented with the number: '1'. The sequence analysis also allows an analysis of the items with different sizes. In this figure, we can also see that not all the items contain both stores and duration. In the third sequence of the figure only the fifth item contained both of these elements. By adding this attribute we can discover similar informations of the previous representation but taking into account the duration.

```
1 -1 1 509 -1 3 536
1 -1 2 -1 3 531 -1 3 536
1 -1 3 -1 1 -1 1 -1 3 531 -1 3 536
```

Figure 3.7: Representation of sequences composed of stores and respective duration of the visit

The next representation is very similar to the previous one. These sequences contain the store's number but instead of the duration, they contain the time of day of that visit. Figure 3.8 represents some of these sequences. The first element of the first line is '10 375', meaning that the customer visited store '375' in the morning: '10'. In this representation, we notice that it is possible that sequences may contain elements of different times of the day. By adding the time of the day, we can discover sequence patterns of stores in specific periods.

```
10 375 -1 10 509 -1 11 381-2
11 370 -1 11 531 -1 -2
11 329 -1 12 330 -1 -2
```

Figure 3.8: Representation of sequences composed of stores and respective time of day of the visit

The last representation combines all information discussed so far. Each element contains the store number, the duration of the visit and the time of the day where that visit occurred. In figure we can see some examples of those sequences. For example, the first element of the first sequence is a long visit to the store number 381 in the early afternoon period.

```
381 3 11 -1 501 3 11 -1 -2
397 2 11 -1 400 2 11 -1 -2
391 0 12 -1 392 0 12 -1 -2
364 0 12 -1 495 0 12 -1 -2
```

Figure 3.9: Representation of sequences composed of stores and respective duration and time of day of the visit

-

Chapter 4

Results

In this chapter, we discuss the results obtained in this project. It is divided into four sections, one for each each type of sequence representation analysed. A final section discusses results in general. The results refer to the analysis of the tests in several parameters. We will not present all the results for each representation but rather examples of information found in those results. In most cases, the tests can be adapted to get more abroad results with the same methods that we have used.

As each of the representations contains different information, the methods also vary. All the tests presented were obtained with multiple test and evaluation cycles, until the amount of desired registers was obtained. The tests only refer to the original dataset. The main goal of this chapter is to answer the questions raised in section 3.3. The meaning of the word pattern in this dissertation is referring to frequent sequences that we founded. We use this terminology to distinguish the frequent sequences, that were the results of the experiences, from the normal sequences that we used to make the analysis.

4.1 Representation 1: stores

This representation is the simplest one because the sequences contain only the stores visited. Being simple does not mean that it is not possible to extract important conclusions. The initial approach was to test several values for the minimum support threshold. A suitable value generates a number of results that is not too small or too large. With a threshold of 0.01 (or 1%) we found 325 frequent sequences with two stores, 146 with three stores and 23 with four stores, resulting in a total of 494 patterns with at least 2 stores. To be part of this group the patterns need to have at least 175 matches in the 17539 visits (approximately 1 %).

The patterns that appear more often are:

- **two stores' patterns**
  - **< 531 - 536 >** 1128 times, 6.6 %
  - **< 487 - 536 >** 1165 times, 6.6 %
  - **< 381 - 509 >** 960 times, 6.6 %
- **three stores' patterns**
  - **< 487 - 531 - 536 >** 446 times, 2.5 %
  - **< 381 - 509 - 536 >** 392 times, 2.2 %
  - **< 381 - 501 - 509 >** 374 times, 2.1 %
- **four stores' patterns**
  - **< 385 - 388 - 391 - 391 >** 209 times, 1.2 %
  - **< 377 - 381 - 501 - 509 >** 199 times, 1.1 %
  - **< 381 - 501 - 509 - 536 >** 198 times, 1.1 %

In the results we can observe that stores 531 and 536 are very related in sequence visits. They are present not only in the two stores' patterns but also in the three stores' patterns. But if we notice the stores 381 and 509 are present in the patterns of size three although they do not have so many occurrences as the previous mentioned stores. There are 33 additional patterns with two stores, appearing at least 500 times(2.9%). Appendix A contains a list that contains all the patterns that were found.

The order of the elements is also considered in the analysis. There are stores that appear more often at the beginning and at the end of the patterns. The is to detect the largest discrepancies between the frequency of a store in the original informational and the frequency of this element in the patterns in a certain position. The top five stores in the beginning and at the end of the patterns are:

- **Beginning**
  - Store **381** - 18 times, 5%
  - Store **361** - 18 times, 5%
  - Store **377** - 18 times, 4%
  - Store **370** - 18 times, 4%
  - Store **371** - 18 times, 4%
- **Ending**
  - Store **536** - 53 times, 16%

- Store **509** - 24 times, 7%
- Store **531** - 22 times, 6%
- Store **487** - 16 times, 4%
- Store **501** - 12 times, 3%

If we compare the frequency of store 536 in the ending position which is 16% while its frequency in the data is of 5.5%, this represents a significant difference. Doing this comparison to each store we were able to find some of the biggest detours and why they happened. By finding this discrepancies in the comparison we can conclude what are the stores that are more present in the patterns.

Since the categories of the stores are available in the data, we can analyse these results according to this information. The most frequent categories founded in the patterns are:

- **Clothing in general** - 122 times, 25%
- **Miscellaneous** - 56 times, 11%
- **Ladieswear** - 52 times, 11%
- **Underwear** - 41 times, 8%
- **Shoe Store** - 40 times, 8%

In this list is also described the number of occurrences for each category and the frequency that it represents. With this list we can, for example, compare the frequency of the category in patterns and the frequency of the category in the data. The difference between the frequency in the patterns (25%) and in the data (13%, it can be found in figure 3.5) is quite large.

We also identified frequent sequences that contained only stores from a single category. There were 9 such categories, among which only the "clothing in general" category had more than 4 occurrences, namely 12, representing 3% of the patterns in that situation. Adding all the occurrences that was composed with items of the same type, we only have 25 in total, making only 5% of the patterns. This result can indicate that in the frequent sequences found in our dataset, there are no great similarities of store's categories between the items of the same pattern.

4.2 Representation 2: stores and duration of visits

In this experiment, we add the duration that a customer spent inside of the stores. The sequence mining algorithm was tested with a support threshold of 1%. The results included many patterns containing only duration (without store ids). Then we refined the results, selecting only the patterns that contained at least a store present in one of their items. We found 349 patterns in these circumstances. The patterns that appeared more often are the ones related to the most frequent store, 536:

- **< short , 536 >** - 3692 times, 30%
- **< short , short, 536 >** - 3094 times, 25%
- **< long , 536 >** - 2380 times, 19%

Here we can observe that this store is frequently preceded by short visits in one and two stores. It also is preceded by long visits a large amount of times.

With this representation, we can also extract other types of results. For example, the sub-patterns **< 381 , short >** is present in 1776 patterns, roughly 15% of all patterns. This means that after a visit to the 381 store the following store is visited for a short period.

In addition to these two types of patterns we can also find others in which one of the elements contains at the same time a store and a duration of a visit. But not all these patterns contain information about patterns. For example, **< {long, 536} >** refers to a single element. That element is composed of two values: long and 536. Such information can be obtained with simpler tools. n the other hand, a pattern such as **< long, {long, 536} >** indeed contains information about a pattern. This pattern is presented in 1265 visits and it shows that after a long visit to a store there is often another long visit to store 536.

Stores 486, 349, 377, 378 and 486 are much more often associated to very short visits. Knowing that people spent less time inside a store we know that the probability of a purchase being made is reduced. Store 501 is more often associated with long visits. The pattern **< long, {long, 501} >** indicates that there is a long visit before a long visit to the store 501. Spending at least 30 minutes in a store significantly increases the chance of buying some product in that place.

The duration of visits of customers to stores is an indicator of the probability of buying something. However, all the conclusions that we can draw must be subject to confirmation by other means (e.g. analysing sales data). All of the patterns that contain several elements are specializations of the shorter ones, making many of the patterns very similar among each other. In appendix B the complete set of results concerning this representation are presented.

4.3 Representation 3: stores and time of the day of the visit

Alternatively to representation 2, representation 3 complements the information about the store with the time of the day replacing the duration of the visit. The goal was to observe if there are any resemblances with the patterns found in the experiments with representation 1. First, we divided the data in five subsets, each one corresponding to one time of the day. We applied the sequence mining algorithm to each one of the subsets of data. Because of the small amount of data in each time of the day, we have selected a value of 0.5% for the support threshold, obtaining the following results:

Time of Day	Number of Visits	Number of patterns
morning	949	39
early afternoon	1022	27
afternoon	931	39
evening	1138	32
night	853	34

Table comparing the number of total visits and the numbers of patterns found in five time of day periods

The table shows that a larger number of visits does not necessarily imply a larger number of patterns. The evening period, which is the one with more visits, is only the fourth regarding the number of patterns found. Maybe this happens because the people who go to a shopping centre in this period are more in a hurry and only want to go to a specific store. We can also observe that the period with the least number of visits was the third one regarding the number of patterns found.

The table below presents the top 5 stores according to the number of times they appear in frequent sequences, for each period of the day. Their frequency in the data is also presented (in between parenthesis). It shows, in some cases, large discrepancies between the two frequency values. Store 494 (highlighted in bold) is the second most frequent in patterns regarding the morning period, although it is only the 26th most frequent in the data. In general, this store has a higher frequency in patterns concerning that period, when compared to other times of the day. We can speculate that this store is a good store to invest with if we want to increase our clients numbers in this period, for example. This kind of pattern occurs with several stores. Store 377, despite being only the 12th most frequent in the original data, is represented in the top 5 of the latest three periods (4th in afternoon period, 3rd in evening period and 3rd at night period), concerning frequency in patterns. By observing these differences we can conclude what are stores that are more popular in patterns of a determined period of the day.

Rank	Morning	Early afternoon	Afternoon	Evening	Night
1	536 (1)	501 (13)	536 (1)	509 (2)	509 (2)
2	**494 (26)**	509 (2)	509 (2)	501 (13)	536 (1)
3	509 (2)	536 (1)	361 (8)	*377 (12)*	377 (12)
4	486 (17)	513 (33)	*377 (12)*	486 (17)	531 (4)
5	385 (7)	391 (27)	531 (4)	391 (27)	501 (13)

Table comparing the Top 5 stores present in the patterns for each time of the day period and the relative position of those stores in the original data between parenthesis

As expected, some stores are frequent both in the data and in the frequent sequences. For instance, Stores 536 (most frequent in the original data) and 509 (second most frequent in the

original data) are present in the top 5 of the different times of the day, in terms of frequency in patterns. On the other hand, store 381, the third most visited store does not appear in the table. That fact is probably due to a uniform distribution of the visits throughout the entire day.

All the patterns founded were composed of two items. For example, in the morning period, one of the patterns found was: **< 531, 536 >**. This pattern indicates that a visit was first made to the store 531 and then to store 536. An analysis made to each pattern could be interesting to a specific store in order to see the comparability of that stores with other in specific times of the day. For example, if we choose the store number 494 we find 5 patterns with it. Being all patterns composed by only two stores, we can speculate that those patterns may indicate five stores that are related with store number 494 in this period.

Other characteristics can be added to the analysis to improve the results. For example, we can also see the most visited categories present in the patterns for each time of day. The table below represents one of the added characteristics. For example, if we look at the morning period, we realize that the children's wear stores are the 2nd most frequently, although it is the 6th most frequent in the original data and it is not present in any other time of the day. The opticians' stores also stand out with the presence in the Top 5 of the 3 later times of day.

Top	Morning: 10 - 13h	Early Afternoon: 13 - 16h	Afternoon: 16 - 19h	Evening: 19 - 22h	Night: 22 - 03h
1	clothing in general (1st)	clothing in general (1st)	clothing in general (1st)	clothing in general (1st)	clothing in general (1st)
2	children's wear (6th)	shoe store in general (8th)	optician (10th)	shoe store in general (8th)	bijouterie (5th)
3	underwear (7th)	miscellaneous (2nd)	miscellaneous (2nd)	jeans and casual wear (11th)	shoe store in general (8th)
4	ladieswear (3rd)	jeans and casual wear (11th)	shoe store in general (8th)	bijouterie (5th)	optician (10th)
5	shoe store in general (8th)	bijouterie (5th)	ladieswear (3th)	optician (10th)	miscellaneous (2nd)

Table comparing the Top 5 stores' categories present in the patterns for each time of the day period and the relative position of those stores' categories in the original data between parenthesis

These results confirm that patterns of visits to stores vary their frequencies depending on the time of the day. However, we note that these results must be further investigated, as the dataset contains few registers of visits in each period. Appendix C contains all the results of these experiments.

4.4 Representation 4: stores, duration and time of the day of the visit

The tests made with this representation were the most complex. In addition to the normal analysis of the previous tests, where we only considered patterns that had at least one store, we also did an analysis of the most common patterns for the most common values (a specific store, which contains 180 possible different values, has much less chance of appearing in a pattern than any reference to the duration of a visit, which contains 4 different values, and the height of the day, which contains 5 different values).

After combining the three parameters, we tested the sequence mining algorithm with several values for the support threshold. With it at 0.1, equivalent to 10%, there are only patterns which repeat the elements in each item, which makes sense, especially on the times of the day. For instance, the **< morning, morning >** and **< short, short >** patterns. By decreasing this value, the number of patterns found increases. For a threshold of 0.08, equivalent to 8%, we detect some interesting patterns like **< evening, {long, evening}>** and **< {long, evening}, evening>**. These patterns indicate that, in the evening, customers tend to include at least one long visit to one store. The study of this kind of patterns can contain important information about the general big picture of the visits in a shopping centre. By analysing the frequent sequences in more high support thresholds we can observe the most frequent sequence patterns and detect some general shopping centre's paths made by the consumers. If we lower the threshold, we find more specific sequences and detect sequence patterns of a more specific parameter like a time of the day period.

In the following tests, we decreased the threshold value until the algorithm found patterns with stores, which happened for a value of 0.01 (1%). However, virtually all of the patterns found were repeated from the ones found in the three previous analyses. This happened because almost every pattern did not contain elements of the time of the day and duration simultaneously.

The first pattern that contained at least one element of the three parameters was present in 0.96% of the patterns. The pattern is **< evening, {long, evening, 536}>**. The meaning of it is that first we have a visit in the evening period, and after that visit we have a long visit, in the evening period, to store 536. This pattern is very similar to the one found in the previous experiment but included a store in the second item. The other patterns that were in the same circumstances have a very small frequency. With a considerably large dataset, even with a small threshold, those patterns could be much more significant because there were many occurrences to corroborate them. This means that further experiments must be carried out to consolidate these results.

Additionally, stores 536 and 509 are the most frequently found in the patterns, confirming their frequency on the original data. Any of the frequent sequences found contained more than one time of the day simultaneously. Many of those contained the same time of day in all elements besides other items. For example: **< {long, evening}, evening, {very short, evening}>**. This

probably happened because the times of the day periods were well divided and they extend for 3 hours each, which gives a much longer time total than most of visits' duration. All the results of this experience can be observed in appendix D.

4.5 Results Discussion

As mentioned above, all the results presented in this chapter are examples that illustrate the kind of information we can obtain in the experiments we made. In most cases, these results were the most significant, i.e., the results that contained the highest number of occurrences. Additional information can be obtained with an analysis of the results in the appendices.

Most of the goals we set out to investigate have been met. The results from the tests proved that it is possible to discover potentially valuable information applied to store visits data in a retail space. Patterns of four different types were found:

- Patterns that only contained stores allowed the extraction of patterns describing frequent patterns of store visits.
- Patterns that contained stores and duration enabled the algorithm to obtain patterns of visits to stores together with an indication of the interest of the store to the customer.
- Patterns that contained stores and times of the day allowed the extraction of patterns separately for different periods of the day.
- Patterns that contained stores, duration and times of the day enabled the extraction of more fine-grained information about customer habits. Through this information we can possibly make a more detailed study to find some costumer behaviour habits.

Data preparation was very important for a much better analysis of the data. From the 45092 visits, only 17539 had at least two stores. There must be at least two elements in a pattern for the results to be interesting. A study of the available dataset should be done so that we can eliminate information that will not be useful to us before starting the experiments.

The analysis carried out here is very broad and the results obtained would be interesting for the owners of a shopping centre and possibly also for the owners of stores. Using the tests carried out, you can pick up a specific store or set of stores and make a more specific analysis of the results as indicated in section 4.3. The patterns could be used to find what sequences of stores their clients visit more frequently; which stores they visit before and after the visit to a particular store; what happens in a specific time of the day; if the duration of a visit to a given store influences the duration of the following visits.

Chapter 5

Conclusions and Future Work

We are at a stage where access to information increases from day to day. The company that provided us with the data has the ability to collect the location of people inside indoor spaces, namely shopping centres. Analysing that type of data can bring us more information about the consumers and their behaviours. The present dissertation discusses the use of sequence mining techniques to identify patterns of customer behaviour in a shopping centre. Prior to this project, the use of sequence mining techniques in the retail data provided by Movvo has not been done. The use of these techniques enabled the extraction of sequential patterns related to the this data.

The dataset provided consists of a collection of visits, each from a person to a particular store at a particular time. Each of these visits contained various types of information. Through this information we were able to analyse different factors: store category, duration of a visit and the time of the day that a visit was made.

The study that we made proposes a new approach to the analysis of the provided dataset. The developed methodology applies sequence mining techniques regardless of the size of the dataset. The use of sequence mining allows discovering several types of patterns with items of different types and sizes. The evaluation of the performed analyses allowed us to confirm the potential that exists in generalizing the use of these techniques in retail related datasets.

The project was divided into four experiments with four different sequence representations. The first representation was composed by stores. The second representation was composed by the stores and the duration of the visit in that store. The third representation was composed by the stores and the time of the day of the visit. The fourth representation was composed with stores, the duration of the visit in that store and the time of the day of the visit. In each of these experiments, we analysed the different representations in the information to obtain different conclusions in each one. In each experience we had to do little adaptations. The adaptations used were also different in each experience and are explained the reasons for that during the dissertation.

We used an implementation of several sequence mining algorithms called SPMF. After several tests, the SPAM algorithm was chosen because it was the most flexible since it allowed the selection of a minimum number of elements in the sequences, and for the efficient implementation. Several smaller programs in the Java language were written to help the conversions of the data in the SPMF desired format.

5.1 Goals achieved

The main goal of detecting different sequence patterns in the data was achieved. In the experiments, we detected patterns of stores, stores and their duration and stores in specific times of the day. To be able to do this we had to adjust the support threshold of the sequence mining to minimise the number of results and select the most important ones. The support threshold is a parameter that corresponds to the minimum amount of occurrences that a pattern need to fulfil to be present in the founded patterns.

It was possible to obtain some patterns that contained stores, times of the day and duration of the visit. However, the available dataset was relatively small (one week of data) which reduces the generality and the strength of the conclusions. To consolidate the results of this work, a much bigger dataset is needed.

The detection of patterns of sequences in the shopping centre was also conducted with success. By analysing the most frequent sequences we were able to observe general patterns of buyers. Those patterns contained information like the duration and the time of the day.

In terms of potential value to business we discover some interesting patterns. We have found general patterns of frequent sequences that can indicate what are the most frequent sequences made by the clients of this shopping centre. This added value can culminate in some decisions in a shopping centre like detect the most popular stores in sequences, detect the patterns in different times of the day, etc. Detecting patterns related to a specific store is also possible with our analysis. We can make a specific analysis of the results obtained to a specific store and discover what are the stores that are more present together in the patterns.

The patterns obtained were also analysed in terms of the categories of the stores. These frequent sequences found contained some repeated categories in its elements but in most of the visits, the categories were not repeated. We could not conclude that there were sequence patterns of buyers looking for a specific product in many stores of the same category. The category topic could be replaced with another topic in order to analyse other types of information. We did not use many topics due to the lack of them in the database.

5.2 Future Work

The dataset we worked with, despite having 100,000 records, was not enough to get the best results in all experiments. It corresponds to a single week in terms of time. A much larger dataset, relative to a longer period is necessary to be able to confirm and improve the results of the experiments we carried out.

A possible integration of some of these analyses into a company's workflow could also be implemented. It would only require an adaptation of the format of the data and also some automation processes such as the use of the algorithm. It could be included in the analysis process already existing. It would however increase the computational time of that process.

It would also be interesting to make experiments to extract patterns allowing gaps between the elements. If we made an analysis with gap with value '1' we would detect all the sequence patterns that contained one item between any other two elements of a pattern. Adding this factor could increase the number of sequence patterns found because we would exclude the items that were present in between any two elements of a pattern and potentially decrease the number of sequence patterns that we not found due to the presence of a very short visit to a store, for example. In addition, some other variables could be added to bring new information about the dataset. For example, we could have a variable that contained if the visit bring any sale to the store and the value of it. These variables could

Through a deeper analysis of the obtained results, it would be possible to make a study of the buyers' behaviour predictions. In order to do that we would need to make some other analysis and not considerate only the sequence mining analysis. For example, we could add the information of what the client really purchased in each visit. We would also could use other clients' specific attributes like the age, gender, monthly income, etc. That study could lead to a bigger understanding of the clients in a retail space.

Besides gaining a better understanding of customer behaviour, the methods used in this project could be used as the basis for a recommendation system. It would recommend stores to customers depending on the current behaviour during a visit. Such a system could also be extended to make product recommendations by taking customer age, gender, and other characteristics into account. This would possibly require a combination of sequence mining with other data mining techniques.

References

[Abb15] Ahmed Abbasi. Predicting Behavior. (june), 2015.

[Agr] Rakesh Agrawal. Fast Algorithms for Mining Association Rules 1 Introduction. pages 1–32.

[AGYF02] Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick. Sequential pattern mining using a bitmap representation. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435, 2002.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, 1215:487–499, 1994.

[Big12] Ibm what is big data: Bring big data to the enterprise. https://www-01.ibm.com/software/in/data/bigdata/, 2012. Accessed: 2016-07-04.

[BL14] Luke Bermingham and Ickjai Lee. Spatio-temporal sequential pattern mining for tourism sciences. In *Procedia Computer Science*, volume 29, pages 379–389, 2014.

[Bre16] Now they want to know about it. http://fortune.com/2016/06/24/brexit-google-trends/, 2016.

[CBDJ09] Guénaël Cabanes, Younès Bennani, and Frédéric Dufau-Joël. Mining customers' spatio-temporal behavior data using topographic unsupervised learning. In *8th International Conference on Machine Learning and Applications, ICMLA 2009*, pages 372–377. IEEE, dec 2009.

[CCS14] Yen Chun Chou, Howard Hao Chun Chuang, and Benjamin B M Shao. The impact of e-retail characteristics on initiating mobile retail services: A modular innovation perspective. *Information and Management*, 53(4):481–492, 2014.

[CdL03] Colleen Collins-dodd and Tara Lindley. Store brands and retail differentiation : the influence of store image and store brand attitude on store own brand perceptions. 10:345–352, 2003.

[CDL08] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of YouTube videos. *IEEE International Workshop on Quality of Service, IWQoS*, pages 229–238, 2008.

[CS15] Longbing Cao and Technology Sydney. Nonoccurring Behavior Analytics: A New Area. 2015.

[CTG12] Chetna Chand, Amit Thakkar, and Amit Ganatra. Sequential Pattern Mining : Survey and Current Research Challenges. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):185–193, 2012.

[CTK13] A Cagri Tolga, Fatih Tuysuz, and Cengiz Kahraman. a Fuzzy Multi-Criteria Decision Analysis Approach for Retail Location Selection. *International Journal of Information Technology & Decision Making*, 12(4):729–755, 2013.

[ENK06] Frank Eichinger, D.D. Nauck, and Frank Klawonn. Sequence mining for customer behaviour predictions in telecommunications. *Proceedings of the Workshop on Practical Data Mining at ECML/PKDD*, pages 3–10, 2006.

[FVLG+16] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. The spmf open-source data mining library version 2. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 36–40. Springer, 2016.

[FY03] Martin Fishbein and Marco C. Yzer. Using Theory to Design Effective Health Behavior Interventions. *Communication Theory*, 13(2):164–183, 2003.

[GH14] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2014.

[GM15a] Ashish Goel and Bhawana Mallick. Customer Purchasing Behavior using Sequential Pattern Mining Technique. *International Journal of Computer Applications*, 119(1):975–8887, 2015.

[GM15b] Ashish Goel and Bhawana Mallick. Customer Purchasing Behavior using Sequential Pattern Mining Technique. *International Journal of Computer Applications*, 119(1):975–8887, 2015.

[HPK06] J. Han, J. Pei, and M. Kamber. *Data Mining: Concepts and Techniques*, volume 3. Elsevier., 3 edition, 2006.

[KCK02] R. J. Kuo, S. C. Chi, and S. S. Kao. A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. *Computers in Industry*, 47(2):199–214, 2002.

[Lie13] Thomas Liebig. Pedestrian Mobility Mining with Movement Pa erns. 2013.

[Lin13] Z Lin. Indoor Location-based Recommender System. 2013.

[LJ12] Alexandros Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[ME10] Nizar R Mabroukeh and C I Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, 43(1):1–41, 2010.

[MSC14] Viktor Mayer-Schönberger and Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. *International Journal of Advertising*, 33(1):181–183, 2014.

REFERENCES

[MSW14] Abhishek Mukherji, Vijay Srinivasan, and Evan Welbourne. Adding intelligence to your mobile device via on-device sequential pattern mining. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, pages 1005–1014, 2014.

[MT03] R. J. T. Morris and B. J. Truskowski. The evolution of storage systems. *IBM Systems Journal*, 42(2):205–217, 2003.

[Muz12] Muhammad Muzammal. Mining Sequential Patterns from Probabilistic Data by Declaration of Authorship. (September), 2012.

[Oli15] Mariana Rafaela Oliveira. Propositional and Relational Approaches to Spatio-Temporal Data Analysis. 2015.

[TL15] Chieh Yuan Tsai and Bo Han Lai. A Location-Item-Time sequential pattern mining algorithm for route recommendation. *Knowledge-Based Systems*, 73:97–110, 2015.

[WWMS15] Aileen P. Wright, Adam T Wright, Allison B. McCoy, and Dean F Sittig. The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53:73–80, 2015.

[XCA05] J Xia, Vic Ciesielski, and Colin Arrowsmith. Data mining of tourists spatio-temporal movement patterns: A case study on Phillip Island. *Proceedings of the Eighth . . .*, pages 1–15, 2005.

[Yan10] Ping Yan. Spatial-Temporal Data Analytics and Consumer Shopping Behavior Modeling. 2010.

[ZB03] Qiankun Zhao and Sourav S Bhowmick. Sequential Pattern Mining: A Survey. *Technical Report, CAIS*, 2003.

REFERENCES

Appendix A

Store's only representation experience

370 371 501 509	371 381 501 509	377 501 509 536
371 377 378 381	377 378 381 509	381 501 509 536
371 377 378 509	377 378 501 509	385 388 391 392
371 377 381 509	377 381 501 509	
371 377 501 509	377 381 509 536	

Figure A.1: Founded sequences of size 4

315 318 533	361 494 536	370 509 536	375 377 536	377 381 501	381 487 536	388 392 513	487 494 531
315 318 319	361 495 536	371 375 377	375 377 378	377 381 509	381 495 536	388 392 509	487 494 536
323 331 332	361 509 536	371 375 381	375 377 381	377 385 509	381 501 536	388 501 509	487 495 536
323 332 333	370 371 536	371 375 509	375 377 501	377 501 536	381 501 509	388 509 513	487 509 536
342 346 349	370 371 377	371 377 536	375 377 509	377 501 509	381 509 536	388 509 536	494 531 536
346 349 485	370 371 381	371 377 378	375 378 381	377 509 536	385 388 513	391 392 513	494 495 531
349 485 486	370 371 501	371 377 381	375 381 536	378 381 536	385 388 536	391 392 509	494 495 536
349 486 536	370 371 509	371 377 501	375 381 385	378 381 501	385 388 391	484 531 536	494 509 536
349 486 487	370 377 536	371 377 509	375 381 501	378 381 509	385 388 392	484 486 536	495 531 536
349 487 536	370 377 378	371 378 381	375 381 509	378 501 509	385 388 509	484 486 487	495 501 536
361 531 536	370 377 381	371 378 501	375 501 509	378 509 536	385 391 392	484 487 536	495 509 536
361 362 536	370 377 501	371 378 509	375 509 536	381 531 536	385 391 509	485 486 536	501 509 536
361 365 536	370 377 509	371 381 536	377 378 536	381 385 536	385 392 509	485 486 487	509 531 536
361 377 536	370 381 536	371 381 501	377 378 379	381 385 388	385 501 536	485 487 536	
361 381 536	370 381 501	371 381 509	377 378 381	381 385 391	385 501 509	486 487 536	
361 381 509	370 381 509	371 385 509	377 378 501	381 385 501	385 509 536	486 487 488	
361 487 531	370 495 536	371 501 536	377 378 509	381 385 509	388 391 513	486 488 536	
361 487 536	370 501 536	371 501 509	377 381 536	381 388 536	388 391 392	487 531 536	
361 494 531	370 501 509	371 509 536	377 381 385	381 388 509	388 391 509	487 488 536	

Figure A.2: Founded sequences of size 3

513 536	319 323	342 485	361 536	370 385	375 381	381 536	388 536	470 473	487 509
524 525	319 329	342 486	361 362	370 388	375 385	381 385	388 391	470 484	488 531
524 536	319 458	343 536	361 365	370 487	375 388	381 386	388 392	470 485	488 536
525 1288	323 533	343 346	361 370	370 494	375 501	381 388	388 395	470 486	493 531
531 536	323 536	343 349	361 371	370 495	375 509	381 391	388 487	470 487	493 536
531 1175	323 329	343 480	361 372	370 501	376 377	381 392	388 495	470 509	494 531
533 536	323 331	343 484	361 377	370 506	377 531	381 470	388 501	480 536	494 536
536 1175	323 332	343 485	361 378	370 509	377 536	381 486	388 509	480 485	494 495
307 524	323 333	343 486	361 381	371 531	377 378	381 487	391 513	480 486	494 501
309 1288	323 458	343 487	361 385	371 536	377 379	381 494	391 536	484 531	494 509
309 310	323 470	346 536	361 388	371 375	377 381	381 495	391 392	484 536	495 531
310 1288	323 509	346 349	361 487	371 377	377 385	381 501	391 509	484 485	495 536
315 531	325 533	346 485	361 494	371 378	377 388	381 506	392 513	484 486	495 501
315 533	325 329	346 486	361 495	371 381	377 392	381 509	392 524	484 487	495 509
315 536	329 533	348 536	361 501	371 385	377 487	381 510	392 536	484 488	501 531
315 318	329 536	348 484	361 509	371 388	377 494	385 513	392 394	484 509	501 536
315 319	329 458	348 486	362 531	371 487	377 495	385 524	392 401	485 531	501 506
315 323	331 332	348 487	362 536	371 494	377 501	385 531	392 509	485 536	501 509
315 329	331 333	349 536	362 365	371 495	377 506	385 536	399 524	485 486	502 536
315 381	332 533	349 352	362 494	371 501	377 507	385 388	399 427	485 487	506 536
315 385	332 536	349 480	365 531	371 506	377 509	385 391	427 524	485 488	506 509
315 388	332 333	349 484	365 536	371 509	378 536	385 392	458 533	486 531	507 536
315 458	332 458	349 485	366 536	372 531	378 379	385 487	458 536	486 536	509 513
315 487	332 468	349 486	366 495	372 536	378 381	385 494	458 468	486 487	509 524
315 509	332 470	349 487	369 536	372 377	378 385	385 495	458 470	486 488	509 531
318 533	333 536	350 536	369 370	372 494	378 501	385 501	467 469	486 509	509 533
318 536	333 470	352 536	370 531	372 495	378 509	385 509	468 533	487 531	509 536
318 319	339 343	352 484	370 536	372 509	379 536	385 510	468 536	487 536	509 510
318 323	339 349	352 485	370 371	373 377	379 381	386 509	468 470	487 1175	
318 329	342 536	352 486	370 375	375 536	379 509	387 388	469 470	487 488	
318 458	342 343	352 487	370 377	375 377	381 513	388 513	470 531	487 494	
319 533	342 346	360 536	370 378	375 378	381 524	388 524	470 533	487 495	
319 536	342 349	361 531	370 381	375 379	381 531	388 531	470 536	487 501	

Figure A.3: Founded sequences of size 2

```
#Estatísticas Gerais
Número de Sequências: 325 + 146 + 23 = 494
Tamanho das Sequências: 2, 3 e 4

#Elementos que aparecem mais freqentemente
loja 536 - 54 + 61 + 3 vezes - 16%
loja 509 - 30 + 46 + 11 vezes - 9%
loja 381 - 26 + 40 + 7 vezes - 8%
loja 531 - 24 vezes - 7%
loja 487 - 24 + 17 vezes - 7%
loja 377 - 22 + 31 + 9 vezes - 6%
loja 385 - 21 + 20 vezes - 6%
loja 370 - 17 + 17 vezes - 5%
loja 501 - 27 + 7 vezes - 18%
loja 371 - 23 + 6 vezes - 15%


#Elementos que aparecem mais vezes no início
18x: 381 ... - 5%
17x: 361 ... - 5%
15x: 377 ... - 4%
15x: 370 ... - 4%
14x: 371 ... - 4%


#Elementos que aparecem mais vezes no fim
53x: ... 536 - 16%
24x: ... 509 - 7%
22x: ... 531 - 6%
16x: ... 487 - 4%
12x: ... 501 - 3%


#Categorias mais frequentes
122x: Categoria CLOTHING IN GENERAL          - 37%
56x: Categoria MISCELLANEOUS                - 17%
52x: Categoria LADIESWEAR                   - 16%
41x: Categoria UNDERWEAR                    - 12%
40x: Categoria SHOE STORE IN GENERAL        - 12%


#Categorias únicas numa sequência
12x: Categoria CLOTHING IN GENERAL          - 3%
4x: Categoria CHILDREN'S WEAR               - 1%
2x: Categoria MISCELLANEOUS                 - 0%
2x: Categoria LADIESWEAR                    - 0%
1x: Categoria VARIOUS DECORATIONS OBJECTS   - 0%
1x: Categoria UNDERWEAR                     - 0%
1x: Categoria STATIONARY & CRAFT ARTICLES   - 0%
1x: Categoria JEWELERY & ACCESSORIES        - 0%
1x: Categoria BIJOUTERIE                    - 0%
```

Figure A.4: General Results

Appendix B

Stores and visit duration representation experience

1 -1 1 -1 1 -1 1 -1 3 531 -1 3 536

1 -1 1 -1 1 -1 1 531 -1 1 536

1 -1 1 -1 1 -1 1 501 -1 1 509

1 -1 1 -1 1 -1 3 -1 3 531 -1 3 536

1 -1 1 -1 1 -1 3 531 -1 3 536

1 -1 1 -1 1 -1 3 487 -1 3 536

1 -1 1 -1 1 -1 3 509 -1 3 536

1 -1 1 -1 1 531 -1 1 536

1 -1 1 -1 1 377 -1 1 378

1 -1 1 -1 1 377 -1 1 378 -1 1

1 -1 1 -1 1 381 -1 1 -1 1 509

1 -1 1 -1 1 381 -1 1 509

1 -1 1 -1 1 485 -1 1 486

1 -1 1 -1 1 487 -1 1 536

1 -1 1 -1 1 501 -1 1 509

1 -1 1 -1 1 509 -1 1 536

1 -1 1 -1 1 509 -1 3 536

1 -1 1 -1 3 -1 1 -1 3 531 -1 3 536

1 -1 1 -1 3 -1 3 -1 3 531 -1 3 536

1 -1 1 -1 3 -1 3 531 -1 3 536

1 -1 1 -1 3 -1 3 487 -1 3 536

1 -1 1 -1 3 531 -1 3 536

1 -1 1 -1 3 487 -1 1 -1 3 536

1 -1 1 -1 3 487 -1 3 -1 3 536

1 -1 1 -1 3 487 -1 3 536

1 -1 1 -1 3 501 -1 3 509

1 -1 1 -1 3 509 -1 3 536

1 -1 1 531 -1 1 536

1 -1 1 531 -1 3 536

1 -1 1 346 -1 1 349

1 -1 1 346 -1 1 349 -1 1

1 -1 1 349 -1 1 485

1 -1 1 349 -1 1 486

1 -1 1 375 -1 1 381

1 -1 1 377 -1 1 -1 1 -1 1 509

1 -1 1 377 -1 1 -1 1 501

1 -1 1 377 -1 1 -1 1 509

1 -1 1 377 -1 1 378

1 -1 1 377 -1 1 378 -1 1

1 -1 1 377 -1 1 378 -1 1 -1 1

1 -1 1 377 -1 1 378 -1 1 -1 1 -1 1

1 -1 1 377 -1 1 381

1 -1 1 377 -1 1 381 -1 1

1 -1 1 377 -1 1 501

1 -1 1 377 -1 1 509

1 -1 1 381 -1 1 -1 1 509

1 -1 1 381 -1 1 536

1 -1 1 381 -1 1 385

1 -1 1 381 -1 1 501

1 -1 1 381 -1 1 509

1 -1 1 385 -1 1 388

1 -1 1 385 -1 1 388 -1 1

1 -1 1 385 -1 1 509

1 -1 1 388 -1 1 391

1 -1 1 388 -1 1 391 -1 1

1 -1 1 391 -1 1 392

1 -1 1 391 -1 1 392 -1 1

1 -1 1 484 -1 1 487

1 -1 1 485 -1 1 486

1 -1 1 486 -1 1 536

1 -1

1 -1 3 -1 3 509 -1 3 536

1 -1 3 531 -1 3 536

1 -1 3 381 -1 1 -1 3 536

1 -1 3 381 -1 1 -1 3 509

1 -1 3 381 -1 3 -1 3 536

1 -1 3 381 -1 3 -1 3 509

1 -1 3 381 -1 3 536

1 -1 3 381 -1 3 509

1 -1 3 385 -1 3 388

1 -1 3 385 -1 3 509

1 -1 3 486 -1 3 536

1 -1 3 487 -1 1 -1 3 536

1 -1 3 487 -1 3 -1 3 536

1 -1 3 487 -1 3 531

1 -1 3 487 -1 3 536

1 -1 3 494 -1 3 536

1 -1 3 495 -1 3 536

1 -1 3 501 -1 3 509

1 -1 3 509 -1 3 536

1 -1 487 -1 3 531 -1 3 536

1 531 -1 1 536

1 531 -1 3 536

1 315 -1 1 318

1 315 -1 1 318 -1 1

1 318 -1 1 -1 1 533

1 318 -1 1 533

1 323 -1 1 331

1 323 -1 1 332

1 323 -1 1 332 -1 1

1 329 -1 1 533

1 331 -1 1 332

1 332 -1 1 333

1 332 -1 1 333 -1 1

1 342 -1 1 346

1 342 -1 1 346 -1 1

1 342 -1 1 349

1 342 -1 1 349 -1 1

1 342 -1 1 349 -1 1 -1 1

1 343 -1 1 349

1 343 -1 1 349 -1 1

1 346 -1 1 349

1 346 -1 1 349 -1 1

1 346 -1 1 349 -1 1 -1 1

1 349 -1 1 -1 1 486

1 349 -1 1 485

1 349 -1 1 486

1 361 -1 1 -1 1 -1 1 536

1 361 -1 1 -1 1 531

1 361 -1 1 -1 1 536

1 361 -1 1 -1 3 536

1 361 -1 1 531

1 361 -1 1 536

1 361 -1 1 365

1 361 -1 1 365 -1 1

1 361 -1 3 536

1 370 -1 1 -1 1 501

1 370 -1 1 536

1 370 -1 1

1 381 -1 1 385

1 381 -1 1 385 -1 1

1 381 -1 1 385 -1 1 -1 1

1 381 -1 1 388

1 381 -1 1 501

1 381 -1 1 501 -1 1

1 381 -1 1 509

1 381 -1 1 509 -1 1

1 381 -1 3 536

1 385 -1 1 -1 1 -1 1 509

1 385 -1 1 -1 1 509

1 385 -1 1 388

1 385 -1 1 388 -1 1

Figure B.1: Founded sequences (part 1)

1 385 -1 1 388 -1 1 -1 1
1 385 -1 1 391
1 385 -1 1 391 -1 1
1 385 -1 1 392
1 385 -1 1 392 -1 1
1 385 -1 1 509
1 385 -1 3 536
1 388 -1 1 -1 1 513
1 388 -1 1 513
1 388 -1 1 391
1 388 -1 1 391 -1 1
1 388 -1 1 391 -1 1 -1 1
1 388 -1 1 392
1 388 -1 1 392 -1 1
1 388 -1 1 509
1 391 -1 1 392
1 391 -1 1 392 -1 1
1 391 -1 1 392 -1 1 -1 1
1 392 -1 1 513
1 392 -1 1 394
1 484 -1 1 486
1 484 -1 1 487
1 484 -1 3 536
1 485 -1 1 486
1 486 -1 1 536
1 486 -1 1 487
1 486 -1 1 488
1 487 -1 1 -1 1 536
1 487 -1 1 531
1 487 -1 1 536
1 487 -1 3 536
1 494 -1 1 531
1 494 -1 1 536
1 494 -1 1 495
1 501 -1 1 536
1 501 -1 1 509
1 501 -1 3 536
1 509 -1 1 536
1 509 -1 3 536
2 -1 1 -1 3 531 -1 3 536
2 -1 3 -1 3 531 -1 3 536
2 -1 3 531 -1 3 536
2 -1 3 487 -1 3 536
2 -1 3 509 -1 3 536
3 -1 1 -1 1 -1 1 -1 3 531 -1 3 536
3 -1 1 -1 1 -1 3 -1 3 531 -1 3 536
3 -1 1 -1 1 -1 3 531 -1 3 536
3 -1 1 -1 1 -1 3 487 -1 3 536
3 -1 1 -1 3 -1 3 -1 3 531 -1 3 536
3 -1 1 -1 3 -1 3 531 -1 3 536
3 -1 1 -1 3 -1 3 487 -1 3 536
3 -1 1 -1 3 531 -1 3 536
3 -1 1 -1 3 487 -1 3 536
3 -1 3 381 -1 3 509
3 -1 3 388 -1 3 509
3 -1 3 487 -1 1 -1 3 536
3 -1 3 487 -1 3 -1 3 536
3 -1 3 487 -1 3 531
3 -1 3 487 -1 3 536
3 -1 3 495 -1 3 536
3 -1 3 501 -1 3 509
3 -1 3 509 -1 3 536
3 -1 487 -1 3 531 -1 3 536
3 531 -1 3 536
3 361 -1 1 -1 1 -1 3 536
3 361 -1 1 -1 3 -1 3 -1 3 536
3 361 -1 1 -1 3 -1 3 536
3 361 -1 3 -1 1 -1 3 -1 3 536
3 361 -1 3 -1 1 -1 3 536
3 361 -1 3 -1 3 -1 1 -1 3 536
3 361 -1 3 -1 3 -1 3 -1 3 536
3 361 -1 3 -1 3 -1 3 536
3 361 -1 3 -1 3 536
3 361 -1 3 536
3 370 -1 3 536
3 371 -1 3 381
3 371 -1 3 509
3 375 -1 3 381
3 377 -1 3 381
3 377 -1 3 509
3 381 -1 1 -1 1 -1 3 536
3 381 -1 1 -1 1 -1 3 509
3 381 -1 1 -1 3 -1 3 536
3 381 -1 1 -1 3 -1 3 509
3 381 -1 1 -1 3 536
3 381 -1 1 -1 3 509
3 381 -1 3 -1 1 -1 3 536
3 381 -1 3 -1 1 -1 3 509
3 381 -1 3 -1 3 -1 3 536
3 381 -1 3 -1 3 -1 3 509
3 381 -1 3 -1 3 536
3 381 -1 3 -1 3 509
3 381 -1 3 536
3 381 -1 3 385
3 381 -1 3 385 -1 1
3 381 -1 3 385 -1 3
3 381 -1 3 501
3 381 -1 3 509
3 381 -1 3 509 -1 3
3 385 -1 1 -1 3 509
3 385 -1 3 -1 3 509
3 385 -1 3 536
3 385 -1 3 388
3 385 -1 3 388 -1 1
3 385 -1 3 388 -1 3
3 385 -1 3 388 -1 3 -1 1
3 385 -1 3 388 -1 3 -1 3
3 385 -1 3 509
3 388 -1 3 509
3 484 -1 3 536
3 484 -1 3 487
3 486 -1 3 536
3 487 -1 1 -1 3 536
3 487 -1 3 -1 3 536
3 487 -1 3 531
3 487 -1 3 536
3 487 -1 531 -1 3 536
3 494 -1 3 536
3 495 -1 3 536
3 501 -1 3 536
3 501 -1 3 509
3 509 -1 3 536
388 -1 1 391 -1 1 392
487 -1 3 531 -1 3 536
510 3 -1 536 1

Figure B.2: Founded sequences (part 2)

```
#Estatísticas Gerais
Número de Sequências: 349
Ficheiro de Entrada: src/outputStoreDuration/itemsStores+duration.txt
Ficheiro de Saída: src/outputStoreDuration/itemsStores+durationAnalysis.txt


#Elementos que aparecem mais freqentemente
loja 1 - 657 vezes - 188%
loja 3 - 483 vezes - 138%
loja 536 - 156 vezes - 44%
loja 509 - 76 vezes - 21%
loja 381 - 67 vezes - 19%
loja 531 - 52 vezes - 14%
loja 487 - 45 vezes - 12%
loja 377 - 39 vezes - 11%
loja 385 - 32 vezes - 9%
loja 501 - 28 vezes - 8%

#Elementos que aparecem mais freqentemente em pares
3 536 - 131 vezes
3 509 - 41  vezes
3 531 - 41  vezes
1 377 - 37  vezes
3 381 - 36  vezes
1 509 - 35  vezes
1 381 - 31  vezes
3 487 - 31  vezes
1 536 - 25  vezes
1 501 - 19  vezes

#Elementos que aparecem mais vezes no início
235x: 1 ... - 67%
106x: 3 ... - 30%
5x: 2 ... - 1%
1x: 510 ... - 0%
1x: 487 ... - 0%


#Categorias mais frequentes
191x: Categoria CLOTHING IN GENERAL          - 54%
79x: Categoria BIJOUTERIE                   - 22%
76x: Categoria JEANS AND CASUAL WEAR        - 21%
67x: Categoria MISCELLANEOUS                - 19%
52x: Categoria LADIESWEAR                   - 14%
```

Figure B.3: General Results

Appendix C

Stores and time of the day representation experience

531 536	346 349	361 531	366 495	377 378	385 513	388 391	494 531
323 331	348 536	361 362	370 509	377 501	385 387	484 536	494 536
342 344	349 352	361 365	372 494	377 507	385 388	485 486	494 495
343 485	349 485	361 372	375 381	381 501	385 509	486 536	501 509
343 486	349 486	362 494	376 377	381 509	388 513	487 536	

Figure C.1: Founded sequences in morning period

531 536	333 470	370 371	375 501	378 501	388 391	484 536
315 318	349 486	370 501	377 378	381 509	391 513	487 536
323 533	352 486	371 509	377 501	385 388	391 392	501 509
323 332	361 531	375 381	377 509	388 513	392 513	

Figure C.2: Founded sequences in early afternoon period

531 536	352 486	361 365	371 509	378 509	388 391	468 470	493 536
315 318	352 487	361 494	376 377	381 509	391 513	484 487	494 536
323 332	359 361	362 494	377 378	385 388	391 392	485 486	501 509
343 480	361 531	365 531	377 501	385 509	392 394	486 536	
349 485	361 536	370 371	377 509	388 513	399 427	487 536	

Figure C.3: Founded sequences in afternoon period

531 536	343 486	361 362	371 509	377 509	385 509	392 513	484 487
315 318	348 484	370 501	376 377	378 379	388 391	395 397	485 486
318 533	349 486	371 381	377 378	381 501	391 513	466 467	487 536
332 333	359 361	371 501	377 495	381 509	391 392	467 469	501 509

Figure C.4: Founded sequences in evening period

531 536	342 486	370 536	377 379	385 388	392 513	487 531
315 318	349 484	370 371	377 501	385 509	392 394	487 536
318 533	361 362	375 381	377 509	388 513	458 533	494 495
332 333	361 495	375 509	381 501	388 391	484 536	501 509
342 344	365 531	377 378	381 509	391 392	486 488	

Figure C.5: Founded sequences at night period

Appendix D

Stores, visit duration and time of the day representation experience

0 -1 0 -1 #SUP: 922
0 -1 0 10 -1 #SUP: 190
0 -1 0 11 -1 #SUP: 211
0 -1 0 12 -1 #SUP: 139
0 -1 0 13 -1 #SUP: 241
0 -1 0 14 -1 #SUP: 141
0 -1 0 536 -1 #SUP: 56
0 -1 0 509 -1 #SUP: 57
0 -1 1 -1 #SUP: 222
0 -1 1 13 -1 #SUP: 58
0 -1 2 -1 #SUP: 138
0 -1 3 -1 #SUP: 179
0 -1 10 -1 #SUP: 286
0 -1 11 -1 #SUP: 322
0 -1 12 -1 #SUP: 233
0 -1 13 -1 #SUP: 372
0 -1 14 -1 #SUP: 222
0 -1 536 -1 #SUP: 101
0 -1 486 -1 #SUP: 51
0 -1 501 -1 #SUP: 63
0 -1 509 -1 #SUP: 107
0 10 -1 #SUP: 374
0 10 -1 0 -1 #SUP: 190
0 10 -1 0 10 -1 #SUP: 190
0 10 -1 10 -1 #SUP: 286
0 11 -1 #SUP: 411
0 11 -1 0 -1 #SUP: 211
0 11 -1 0 11 -1 #SUP: 211
0 11 -1 11 -1 #SUP: 322
0 12 -1 #SUP: 308
0 12 -1 0 -1 #SUP: 140
0 12 -1 0 12 -1 #SUP: 139
0 12 -1 12 -1 #SUP: 233
0 13 -1 #SUP: 450
0 13 -1 0 -1 #SUP: 241
0 13 -1 0 13 -1 #SUP: 240
0 13 -1 1 -1 #SUP: 58
0 13 -1 1 13 -1 #SUP: 58
0 13 -1 13 -1 #SUP: 371
0 14 -1 #SUP: 299
0 14 -1 0 -1 #SUP: 140
0 14 -1 0 14 -1 #SUP: 140
0 14 -1 14 -1 #SUP: 221
0 536 -1 #SUP: 85
0 349 -1 #SUP: 52
0 361 -1 #SUP: 58
0 377 -1 #SUP: 77
0 378 -1 #SUP: 54
0 381 -1 #SUP: 64
0 388 -1 #SUP: 50
0 486 -1 #SUP: 54
0 487 -1 #SUP: 53
0 501 -1 #SUP: 61
0 509 -1 #SUP: 92
1 -1 0 -1 #SUP: 184
1 -1 1 -1 #SUP: 458
1 -1 1 10 -1 #SUP: 106
1 -1 1 11 -1 #SUP: 97
1 -1 1 12 -1 #SUP: 84
1 -1 1 13 -1 #SUP: 107
1 -1 1 14 -1 #SUP: 64
1 -1 2 -1 #SUP: 133
1 -1 3 -1 #SUP: 103
1 -1 10 -1 #SUP: 195
1 -1 11 -1 #SUP: 162
1 -1 12 -1 #SUP: 158
1 -1 13 -1 #SUP: 199
1 -1 14 -1 #SUP: 140
1 -1 536 -1 #SUP: 69
1 10 -1 #SUP: 259
1 10 -1 1 -1 #SUP: 106
1 10 -1 1 10 -1 #SUP: 106
1 10 -1 10 -1 #SUP: 195
1 11 -1 #SUP: 252
1 11 -1 1 -1 #SUP: 97
1 11 -1 1 11 -1 #SUP: 97
1 11 -1 11 -1 #SUP: 162
1 12 -1 #SUP: 248
1 12 -1 1 -1 #SUP: 84
1 12 -1 1 12 -1 #SUP: 84
1 12 -1 12 -1 #SUP: 158
1 13 -1 #SUP: 300
1 13 -1 1 -1 #SUP: 107
1 13 -1 1 13 -1 #SUP: 107
1 13 -1 13 -1 #SUP: 199
1 14 -1 #SUP: 211
1 14 -1 1 -1 #SUP: 64
1 14 -1 1 14 -1 #SUP: 64
1 14 -1 14 -1 #SUP: 140
1 536 -1 #SUP: 57
1 509 -1 #SUP: 51
2 -1 0 -1 #SUP: 115
2 -1 1 -1 #SUP: 155
2 -1 2 -1 #SUP: 658
2 -1 2 10 -1 #SUP: 113
2 -1 2 11 -1 #SUP: 145
2 -1 2 12 -1 #SUP: 97
2 -1 2 13 -1 #SUP: 175
2 -1 2 14 -1 #SUP: 128
2 -1 3 -1 #SUP: 219
2 -1 3 10 -1 #SUP: 51
2 -1 3 13 -1 #SUP: 65
2 -1 10 -1 #SUP: 201
2 -1 11 -1 #SUP: 244
2 -1 12 -1 #SUP: 181
2 -1 13 -1 #SUP: 284
2 -1 14 -1 #SUP: 199
2 -1 536 -1 #SUP: 79
2 -1 509 -1 #SUP: 83
2 10 -1 #SUP: 281
2 10 -1 2 -1 #SUP: 114
2 10 -1 2 10 -1 #SUP: 113
2 10 -1 3 -1 #SUP: 51
2 10 -1 3 10 -1 #SUP: 51
2 10 -1 10 -1 #SUP: 201
2 11 -1 #SUP: 308
2 11 -1 2 -1 #SUP: 145
2 11 -1 2 11 -1 #SUP: 145
2 11 -1 11 -1 #SUP: 244
2 12 -1 #SUP: 256
2 12 -1 2 -1 #SUP: 97
2 12 -1 2 12 -1 #SUP: 97
2 12 -1 12 -1 #SUP: 181
2 13 -1 #SUP: 393
2 13 -1 2 -1 #SUP: 175
2 13 -1 2 13 -1 #SUP: 175
2 13 -1 3 -1 #SUP: 65
2 13 -1 3 13 -1 #SUP: 65
2 13 -1 13 -1 #SUP: 284
2 14 -1 #SUP: 268
2 14 -1 2 -1 #SUP: 128
2 14 -1 2 14 -1 #SUP: 128
2 14 -1 14 -1 #SUP: 199
2 536 -1 #SUP: 76
2 377 -1 #SUP: 59
2 487 -1 #SUP: 52
2 501 -1 #SUP: 54
2 509 -1 #SUP: 88
3 -1 0 -1 #SUP: 147
3 -1 1 -1 #SUP: 75
3 -1 513 -1 #SUP: 53
3 -1 2 -1 #SUP: 176
3 -1 3 -1 #SUP: 1345
3 -1 3 -1 3 -1 #SUP: 100
3 -1 3 10 -1 #SUP: 188
3 -1 3 11 -1 #SUP: 284
3 -1 3 12 -1 #SUP: 223
3 -1 3 13 -1 #SUP: 319

Figure D.1: Founded sequences (part 1)

3 -1 3 14 -1 #SUP: 331
3 -1 3 536 -1 #SUP: 145
3 -1 3 509 -1 #SUP: 97
3 -1 10 -1 #SUP: 249
3 -1 11 -1 #SUP: 358
3 -1 12 -1 #SUP: 286
3 -1 13 -1 #SUP: 405
3 -1 14 -1 #SUP: 401
3 -1 531 -1 #SUP: 54
3 -1 536 -1 #SUP: 175
3 -1 487 -1 #SUP: 52
3 -1 494 -1 #SUP: 52
3 -1 501 -1 #SUP: 62
3 -1 509 -1 #SUP: 122
3 513 -1 #SUP: 55
3 10 -1 #SUP: 337
3 10 -1 3 -1 #SUP: 188
3 10 -1 3 10 -1 #SUP: 188
3 10 -1 10 -1 #SUP: 249
3 11 -1 #SUP: 454
3 11 -1 3 -1 #SUP: 284
3 11 -1 3 11 -1 #SUP: 284
3 11 -1 11 -1 #SUP: 357
3 12 -1 #SUP: 351
3 12 -1 3 -1 #SUP: 223
3 12 -1 3 12 -1 #SUP: 223
3 12 -1 12 -1 #SUP: 286
3 13 -1 #SUP: 524
3 13 -1 3 -1 #SUP: 319
3 13 -1 3 13 -1 #SUP: 319
3 13 -1 13 -1 #SUP: 405
3 13 536 -1 #SUP: 51
3 14 -1 #SUP: 480
3 14 -1 3 -1 #SUP: 331
3 14 -1 3 14 -1 #SUP: 331
3 14 -1 14 -1 #SUP: 401
3 531 -1 #SUP: 93
3 536 -1 #SUP: 216
3 343 -1 #SUP: 52
3 349 -1 #SUP: 56
3 361 -1 #SUP: 94
3 361 -1 3 -1 #SUP: 62
3 362 -1 #SUP: 50
3 370 -1 #SUP: 55
3 371 -1 #SUP: 72
3 371 -1 3 -1 #SUP: 51
3 377 -1 #SUP: 105
3 377 -1 3 -1 #SUP: 64
3 381 -1 #SUP: 89
3 385 -1 #SUP: 86
3 385 -1 3 -1 #SUP: 55
3 388 -1 #SUP: 76
3 391 -1 #SUP: 60
3 484 -1 #SUP: 60
3 486 -1 #SUP: 79
3 487 -1 #SUP: 115
3 494 -1 #SUP: 83
3 495 -1 #SUP: 68
3 501 -1 #SUP: 96
3 509 -1 #SUP: 149
10 -1 0 -1 #SUP: 281
10 -1 0 10 -1 #SUP: 281
10 -1 1 -1 #SUP: 179
10 -1 1 10 -1 #SUP: 179
10 -1 2 -1 #SUP: 197
10 -1 2 10 -1 #SUP: 197
10 -1 3 -1 #SUP: 279
10 -1 3 10 -1 #SUP: 279
10 -1 10 -1 #SUP: 905
10 -1 10 -1 10 -1 #SUP: 73
10 -1 10 536 -1 #SUP: 80
10 -1 536 -1 #SUP: 80
10 536 -1 #SUP: 82
10 361 -1 #SUP: 51
10 377 -1 #SUP: 51
10 509 -1 #SUP: 54
11 -1 0 -1 #SUP: 302
11 -1 0 11 -1 #SUP: 302
11 -1 1 -1 #SUP: 188
11 -1 1 11 -1 #SUP: 188
11 -1 2 -1 #SUP: 211
11 -1 2 11 -1 #SUP: 211
11 -1 3 -1 #SUP: 384
11 -1 3 11 -1 #SUP: 384
11 -1 11 -1 #SUP: 1066
11 -1 11 -1 11 -1 #SUP: 72
11 -1 11 536 -1 #SUP: 85
11 -1 11 509 -1 #SUP: 72
11 -1 536 -1 #SUP: 85
11 -1 509 -1 #SUP: 72
11 536 -1 #SUP: 91
11 377 -1 #SUP: 62
11 377 -1 11 -1 #SUP: 54
11 388 -1 #SUP: 52
11 487 -1 #SUP: 57
11 501 -1 #SUP: 58
11 509 -1 #SUP: 80
12 -1 0 -1 #SUP: 219
12 -1 0 12 -1 #SUP: 218
12 -1 1 -1 #SUP: 176
12 -1 1 12 -1 #SUP: 176
12 -1 2 -1 #SUP: 177
12 -1 2 12 -1 #SUP: 177
12 -1 3 -1 #SUP: 288
12 -1 3 12 -1 #SUP: 288
12 -1 12 -1 #SUP: 837
12 -1 12 -1 12 -1 #SUP: 59
12 -1 12 536 -1 #SUP: 69
12 -1 12 509 -1 #SUP: 52
12 -1 536 -1 #SUP: 69
12 -1 509 -1 #SUP: 52
12 536 -1 #SUP: 71
12 509 -1 #SUP: 58
13 -1 0 -1 #SUP: 323
13 -1 0 13 -1 #SUP: 322
13 -1 1 -1 #SUP: 212
13 -1 1 13 -1 #SUP: 212
13 -1 2 -1 #SUP: 289
13 -1 2 13 -1 #SUP: 289
13 -1 3 -1 #SUP: 441
13 -1 3 13 -1 #SUP: 441
13 -1 13 -1 #SUP: 1232
13 -1 13 -1 13 -1 #SUP: 79
13 -1 13 536 -1 #SUP: 103
13 -1 13 509 -1 #SUP: 92
13 -1 536 -1 #SUP: 103
13 -1 509 -1 #SUP: 92
13 531 -1 #SUP: 58
13 536 -1 #SUP: 107
13 361 -1 #SUP: 56
13 377 -1 #SUP: 80
13 377 -1 13 -1 #SUP: 58
13 381 -1 #SUP: 55
13 486 -1 #SUP: 55
13 487 -1 #SUP: 70
13 501 -1 #SUP: 73
13 509 -1 #SUP: 97
14 -1 0 -1 #SUP: 218
14 -1 0 14 -1 #SUP: 218
14 -1 1 -1 #SUP: 137
14 -1 1 14 -1 #SUP: 137
14 -1 2 -1 #SUP: 200
14 -1 2 14 -1 #SUP: 200

Figure D.2: Founded sequences (part 2)

14 -1 3 -1 #SUP: 411

14 -1 3 14 -1 #SUP: 411

14 -1 14 -1 #SUP: 942

14 -1 14 -1 14 -1 #SUP: 80

14 -1 14 536 -1 #SUP: 78

14 -1 14 509 -1 #SUP: 84

14 -1 536 -1 #SUP: 78

14 -1 509 -1 #SUP: 84

14 536 -1 #SUP: 83

14 381 -1 #SUP: 55

14 487 -1 #SUP: 53

14 509 -1 #SUP: 91

531 -1 536 -1 #SUP: 76

361 -1 0 -1 #SUP: 65

361 -1 3 -1 #SUP: 77

370 -1 3 -1 #SUP: 57

371 -1 3 -1 #SUP: 64

377 -1 0 -1 #SUP: 65

377 -1 2 -1 #SUP: 50

377 -1 3 -1 #SUP: 79

377 -1 11 -1 #SUP: 54

377 -1 13 -1 #SUP: 58

377 -1 378 -1 #SUP: 69

381 -1 3 -1 #SUP: 66

381 -1 509 -1 #SUP: 62

385 -1 3 -1 #SUP: 71

388 -1 3 -1 #SUP: 51

487 -1 3 -1 #SUP: 60

487 -1 536 -1 #SUP: 74

501 -1 509 -1 #SUP: 67

Figure D.3: Founded sequences (part 3)